An exploration of the effects of group summative assessment marking on higher education students' overall marks

Submitted for the degree of Doctor of Philosophy
School of Education, University of Durham

Richard James Almond, 2012.

Abstract

Groupwork and group summative assessment (GSA) are important learning, teaching and assessment methods used by many educational institutions, not just universities. The differences between the marks that HEI students were awarded for their own independent individual summative assessment (IISA) work and their GSA marks were explored.

The study topic presented itself while the author was contemplating studying for a first degree, when it became apparent that group working and group summative assessment was included in summative assessment methods used in the chosen programme.

Three data sources were from UK undergraduates and graduates, and one was from Australian PG students. Module marks data were collected from over 4000 HE students. They were divided into eighteen faculty/year data sets from four HEI sources.

A systematic difference was found between the distributions of GSA and IISA marks, supporting Lejk et al. (1999). Lower IISA ability students scored higher in GSA modules than in IISA modules. Higher IISA ability students scored lower in GSA modules.

In addition, the mean GSA mark was higher than the mean IISA mark. The standard deviation of the GSA marks was lower than the SD of the IISA marks. Both of these findings support Downie (2001). The relationship was found to vary between the data sets, modules, assessment items and especially between faculties.

The results and conclusions from this study will empower stakeholders, enabling them to be better informed in their choice of first-degree study programmes. They will also allow the use and impact of GSA to be more transparent and better understood, leading to further research and improvement in practice.

Table of contents

List of figures

List of tables

Abbreviations and glossary

AERA	American Educational Research Association
APA	American Psychological Association
BCE	Before Current Era
BCS	British Computer Society
BSc.	Bachelor of Science, first-degree three-years full-time study or equivalent
CEM	Centre for Evaluation and Monitoring, Durham University (http://www.cemcentre.org/)
CI	Confidence Interval
CIEA	Chartered Institute of Educational Assessors
DCSF	UK government Department for Children, Schools and Families
DfES	UK government Department for Education and Skills
DIF	Differential item functioning
DIUS	UK government Department for Innovation, Universities and Skills
ESRC	Economic and Social Research Council
ERASMUS	European Region Action Scheme for the Mobility of University Students
FE	Further Education
GPA	Grade Point Average (United States of America student summative assessment model)
GSA	Group, or groupwork, summative assessment
HE	Higher Education
HEA	The Higher Education Academy
HEAR	Higher Education Achievement Report, (sometimes referred to as a Transcript)
HEI	Higher Education Institution
HNC	Higher National Certificate
HND	Higher National Diploma
IISA	Independent individual summative assessment, e.g. traditional written, timed, unseen examinations
ITC	Information technology and communications
JISC	Joint Information Systems Committee
LEA	Local Education Authority
MBA	Master of Business Administration
MBTI	Myers-Briggs Type Indicator
MLM	Multilevel modelling
NCEP	U.S. Department of Education National Center for Education Statistics
NCME	National Council on Measurement in Education
NPEC	US Government Department of Education's National Postsecondary Education Cooperative
MSc.	Master of Science post-graduate degree
MSci	Four-year (full-time equivalent) integrated Masters' first-degree
NVQ	National Vocational Qualification
OED	Oxford English Dictionary
OfSTED	UK government Office for Standards in Education
PG	Post Graduate Student, Module or Programme
QAA	Quality Assurance Agency
QCA	UK Qualifications and Curriculum Authority
RAE	The UK government Research Assessment Exercise (RAE) is carried out every 5 years on behalf of the separate UK HE funding councils for England, Scotland, Northern Ireland and Wales. The aim is to evaluate the quality of the research undertaken by the HEIs in those areas. Also, see REF.
RCT	Randomised Control Trial
REF	Research excellence framework will replace RAE in 2014.
RSA	Royal Society for the Encouragement of Arts, Manufacturers and Commerce
RSS feed	This software allows automatic advance warning of journal articles on selected topics to be sent to a personal computer over the internet. One version of the acronym is Really Simple Syndication.
SD	Standard Deviation
SE	Standard Error
TRAMSS	Teaching Resources and Materials for Social Scientists
USA and US	United States of America
Weltanschauung	Personal view of how the world works
YTS	Youth Training Scheme

Declaration

I confirm that the material in this thesis is solely the work of the author and has not previously been submitted for a degree in this or any other university.

Signed … … … … … … … … … … … … … … … … … … Date … … … … … … … … … …

Richard James Almond

Statement of copyright

Acknowledgements

Professor P B Tymms, Head of the School of Education and Director of the Centre for Evaluation and Monitoring, University of Durham;

Professor E L Burd, Department of Computer Science and Deputy Dean of the Graduate School, University of Durham;

All those, both known and unknown, who supplied the data;

John Dewey:

> *"all thinking is research and all research is native, original with him who carries it on, even if everybody else in the world already is sure of what he is still looking for"*

(1916: Chapter XI, Experience and Thinking: 2. Reflection in Experience);

Val, *'Thou wert my guide, philosopher and friend'*, thanks for the t-shirt, the pizzas and everything.

Chapter 1. Study introduction

This study explored the quantitative impact of group summative assessment (GSA) marking on higher education (HE) students overall marks compared to what they might expect from their own independent individual assessment (IISA) marks. It used secondary data.

Previously the question had been asked, *"Are group assignments a legitimate form of assessment"* (Morris and Hayes 1997). Other researchers have noted, *wryly*, according to Strauss (2001:64), that *"There is a strong possibility that some students simply learn most effectively alone"* (Watson and Marshall 1995b). These authors summarise the motivation for this study. This thesis is a response to the question from the former author, and the observation from the latter.

Different methods of marking, i.e. individual or group, may be operationalizing different concepts. Different measures from the operationalization could include for example the student individual ability and/or their attainment in the topic (IISA). This is a typical assumption made for assessment items other than those from GSA coursework. In addition to this, GSA items could also include some measure of candidate individual ability, alongside their groupwork skills. For example, they may be attempting to measure individual group-dynamics skills, or the interactions of the group members in the GSA module.

As will be discussed later, groupworking and GSA are learning, teaching and assessment tools used throughout both academic and vocational formal education. In this thesis, the word *vocational* refers to profession or trade, rather than a person's choice of life career or their predisposition to a particular calling. Group working and GSA are used in compulsory schooling, tertiary and FE (further education) and HE undergraduate and taught postgraduate programmes.

Section 1.1 of this thesis describes the research problem. Section 1.2 outlines the study genesis. The thesis organization is outlined in section 1.3.

1.1 The research problem

This study was an exploration of the quantitative impact of GSA on students overall marks, compared to that of IISA. The research problem was whether GSA made an inappropriate difference and biased a student's overall mark, and hence had an impact on their degree classification award. More especially, but not exclusively as will be explained below, the research interest focused on the effect on undergraduate marks.

The null hypothesis H_0 is that the two summative assessment methods, IISA and GSA, have the same impact on students overall marks. The alternative hypothesis H_1 (also, see Table 1) is that the two summative assessment methods, IISA and GSA, each have a different impact on students overall marks.

The study focused on the effect on HE students marks of their being summatively assessed as a group. The expanded research questions and hypotheses, with the method of analysis, are in Table 1.

Table 1: Research Questions, Hypothesis and Methods of Analysis

Research Question	Hypotheses	Methods of Analysis
Does GSA bias student overall marks?	$H_{0:}$ The two summative assessment methods have the same impact on students' overall marks.	
How does GSA affect student overall marks? How do students IISA marks compare to their GSA marks? What is the relationship between IISA and GSA marks? How does the relationship between IISA and GSA vary? What is the correlation between IISA and GSA marks? How do the regression slopes of the IISA and GSA marks compare when regressed on the IISA rank?	$H_{1:}$ The two summative assessment methods each have a different impact on students overall marks.	Regression, Meta-analysis, Multilevel Modeling, Correlation.
Will uncorrelated data pairs show the same bisection pattern in a dual regression line chart as the study data subjects raw data?	$H_{2:}$ Uncorrelated data will show the same scattergram pattern in a dual-line regression chart as the study data.	Simulation, Correlation
Do IISA and GSA operationalize different concepts, or is one or both unreliable?	$H_{3:}$ There will be a low correlation between IISA and GSA category marks.	Correlation reliability estimates

An important part of the research problem was that in degree awarding higher education institutions (HEIs) that practice GSA, an individual degree classification depends partly on the effort, motivation, skill and ability, of the students peers. Outside their common study modules teaching time, some students in the group might be unknown to the others. On the other hand, some or all of the other group members could also be part of stable, previously established, current friendship, sporting, recreational or study groups, or even of couples in an intimate relationship. In some circumstances, clearly, members of groups allocated ostensibly by the same method, e.g.

self-selection, have had different group experiences, i.e. treatment, prior to forming their group.

This form of summative assessment might have other effects on the student. These could be both quantitative and qualitative. The qualitative effects could include emotional, psychological, financial, or even spiritual issues. This study was not aimed at probing any of these. Neither was this an investigation into the effect of GSA marking in any philosophical or metaphysical manner, nor did it look at the impact on society, although, as stakeholders, GSA does affect society as a whole. Nevertheless, future studies of these effects could yield additional meaningful results leading to further insights into the impact of GSA. In the circumstances, a subtitle to this study might be '*A post-positivist view of how groupwork summative assessment impacts quantitatively on students overall summative assessment marks'.*

This study explored the effects of IISA and GSA marking on degree programmes and modules that used both assessment methods. The extent and direction of any systematic differences between them was also explored. In retrospect, the original research question, hinting at a single number simple answer, was somewhat naïve.

1.2 Genesis

The research question concerned the impact that students group summative assessment marks might have on their overall mark. Did GSA marking have any affect at all? How did marks from a GSA module compare to what students might expect from their individual effort and ability?

It is clear that any discussion on the concept of undergraduate groupwork and of the issues surrounding it is meaningless without a baseline. This baseline forms a foundation of understanding on which such a discussion may proceed. This study provides an understanding of the quantitative effect that GSA might have on the overall individual student's marks and consequently on any similar impact it might have on their degree classification.

In order to improve anything it should first be understood. This is also true of the quantitative impact of GSA. Part of this process is to understand the effects of procedures producing it. The first stage is to learn *'how'*, and then the *'why'*. Only then can suggestions be made for the way in which the rules might be changed to improve practice. Yet, for example, as will be discussed later, the educational research literature argues at length the strengths and weaknesses of the andragogical (see section 2.3.1) and employability issues of GSA, apparently with little if any such

understanding.

1.3 Thesis organisation

In this thesis, the phrase *group assessment* means the same as *group summative assessment* (or GSA). Formative group assessment had no substantive part in this study. Additionally, older texts included in the literature review inevitably quote the masculine pronoun, instead of the current convention of gender neutrality. Such was the style at the time of their publication. There are eight chapters in this thesis. References and appendixes follow. This study cannot follow an existing theoretical framework. There are no published, named theories on the effects of GSA marking. Chapter 2 therefore outlines a conceptual framework for this study rather than a theoretical perspective.

Section 2.1 discusses who the principal stakeholders of GSA are. Section 2.2 concerns theoretical perspective and value neutrality. Section 2.3 is mainly a personal position statement on the concept of an academic first-degree. It also contains a short section, 2.3.1, entitled Andragogy - not pedagogy Section 2.4 explains how the terms *programme* and *module* are used in this thesis. Section 2.5 explains the assumptions made in this study about the word *ability*, the use of scatter plots in this thesis and the use of the judgemental terms *assessment* and *evaluation*. Section 2.6 introduces the main problems associated with GSA. The penultimate section in the second chapter, 2.7, explains the reasons for studying the quantitative impact of GSA marking rather than a qualitative or mixed methods study. This includes the author's work, academic study and research background. Section 2.8 is the chapter summary.

Only a handful of research publications have been found that directly address the focus of this study. The literature review chapters reflect this. Chapter 3 is the first of two literature review chapters. The general background literature to this study is reviewed. The main issues surrounding GSA begins with the difference noted in the literature between the terms *group* and *team* (section 3.1). Groupwork, group work, group and peer learning, and group assessment are reviewed in section 3.2. This is followed by a review of the literature on the practise effect (section 3.3), and of the ubiquitous nature of GSA practice in higher education (section 3.4). This chapter also describes and discusses research literature relating to the rationale, that is, the perceived advantages, for using GSA (section 3.5).

Methods of forming student GSA groups (section 3.6) including the impact of heterogeneous and

homogenous personality groups in section 3.6.3.3, are reviewed, and the limited literature on the impact of GSA is presented and reviewed (section 3.7). An unverified claim for the effect of GSA marking is reviewed in section 3.8. Disadvantages associated with the use of GSA, from the literature, are reviewed in section 3.9. (The advantages are reviewed in section 3.5, the Rationale for practice section.)

The views of some GSA students, staff and alumni are included in section 3.10. There is also a review of the limited research on the degree classification system in section 3.11. Southampton University's *Jumpstart* scheme, that introduces students to groupwork, is reviewed in section 3.12. Section 3.13 is a short review of the limited research of the biographical predictors of student GSA outcomes. The penultimate section is a review of the work referencing andragogy, in work concerning HE students (section 3.14), the preferred term being pedagogy. Finally, in this chapter, section 3.15 summarises the general background literature review.

Chapter 4 reviews the HE summative assessment literature. After an introductory section, it begins with a section reviewing the literature examining the purposes of HE (section 4.2). This is followed by a review of the roles and purposes of assessment in HE (section 4.3). The concept of summative assessment is reviewed in section 4.4. Assessment practice as it applies to GSA is reviewed in section 4.5. It includes a review of the literature on peer- and self-assessment (section 4.5.2), and methods of deriving students marks from group marks (section 4.5.5). Also included in this chapter is a section on summative assessment as a motivator and primary focus for students study regimens (section 4.5.6). Reliability and validity of HE assessment is section 4.6. Portfolios are reviewed in section 4.7. There is also a section on the Oppenheim et al. *Assumptions* that are of relevance to this study topic (section 4.8). One of their main points was that they expected students to take individual responsibility for their work (Oppenheim et al. 1967:341). The *'Standards'* (Baker et al. 1999) views on fairness are reviewed in section 4.9. The penultimate section in chapter 4 reviews the importance of summative assessment in HE (section 4.10). The final part in the chapter is a summary (section 4.11).

Chapter 5 is the Study design considerations: Method and methodology chapter. The data sources are introduced in section 5.1. The hierarchical nature of educational data structure and multilevel modelling are outlined in section 5.2. Consideration of the simulation exercise is found in section 5.3, while extraneous, mediating and intervening variables are considered in section 5.4.

Examples of HEI regulations surrounding groupwork and GSA practice are given in section 5.5. This chapter is summarised in section 5.7. Chapter 6 is the Data description and collection chapter. It includes a section on why undergraduate data dominates this study (section 6.1), and one on ethics (section 6.2). The opportunity sample nature of the data and a description of them and their main collection issues are in section 6.3. There is also a summary (section 6.4).

Chapter 7 is the Data processing, analysis and results chapter. Data from the four separate sources are analysed in section 7.2. Correlations are considered in section 7.3. The regression analysis section is located in section 7.4. Simulation (section 7.7), meta-analysis (section 7.6), multilevel modelling (section 7.5) and a summary section (section 7.8) complete the data analysis chapter. Chapter 8 consists of two discussion sections, the first on the general background literature (section 8.1); the second discusses the HE assessment literature review (8.2). It includes issues that arose during the study, where including them in other sections would not have been appropriate. These include, for example, actual and speculative critical stakeholders' perspectives on GSA practice (section 8.2.12). The study limitations are discussed in section 8.2.14. The study conclusions are in section 8.2.16, recommendations for practice are in section 8.2.17 and some suggestions for topics for further research are in section 8.2.18.

Chapter 2. Conceptual framework

The previous chapter introduced the study. This conceptual framework chapter is an account of the researcher's view of the research problem in relation to his perception of how the world works. After this introductory section, the following sections (beginning with section 2.1), will explain the author's conceptual framework for this study.

This is not a philosophical framework in the Crotty (1998) style, which implied that all the attribute variables are due equal consideration. This could include their many covariances and the even more numerous side issues that this would raise. They may all affect student marks. Many of these, such as their ethnic origin, gender, work and study habits and experience, how their views of their studies might have changed over their duration, and personal interests are unknowable in a secondary data survey study such as this one. They could all have an impact on the effect of GSA.

This chapter is neither a theoretical perspective, nor a theoretical framework. It is a conceptual framework. This is because, as mentioned in Chapter 1, there is not yet any published theory on the effect of GSA marking on students overall marks to guide this exploration.

It has been suggested that traditional written assessments do produce reliable and consistent individual student scores (Huot 1996). This, Huot asserted, was due to their design. He also posited that these traditional practices were *"based upon classical test theory".* Their roots being in a positivist epistemology that assumes *"that there exists a reality out there, driven by immutable natural laws"* (Huot 1996:549; Elton and Johnston 2002:6). Huot also observed that there was an assumption that

> *"student ability in writing, as in anything else, is a fixed, consistent, and a contextual human trait. Our ability to measure such a trait would need to recognize these consistencies and could be built upon psychometrics".*

(Huot 1996:549-550)

Other researchers have made similar assumptions *"that examination results should be distributed in a certain way"* (Oppenheim et al. 1967:349), see section 3.11.

In spite of this, this study was not theory confirming, neither was it theory seeking. It was an attempt to throw some light on the quantitative impact that GSA might have on students overall marks, compared to those from their independent individual efforts. It was an exploration. There is however, an understanding implied by for example the work of Lejk et al. (1999), Hoffman and

Rogelberg (2001) and Almond (2009). In addition, there are general theories and views of assessment. These include item response theory, including a sub-group of it known as Rasch modelling. They also include classical test theory, mentioned by Huot (1996) and cited earlier, which is often associated with the view that assessment marks will be normally distributed. General theories on assessment also include Messick's views on assessment validity (Messick 1980, 1993). These all have a bearing on this study.

Part of this conceptual framework is that this study was an exploration. The results offer an insight into the quantitative impact that GSA has on student overall marks. It is a further step in a research thread, which, it is hoped, will lead to more alternative hypotheses and in particular theories of the practice of GSA. This will lead, in turn, to changes in practice, in how its results are promulgated and in how they are interpreted and understood by stakeholders.

A conceptual framework model has been described and explained by other researchers. They referred to it as *"intermediate theory"* (Shields and Tajalli 2006:313). They explained that it was an idea grounded in the researcher's understanding of the concept. It was not an academically accepted theory. It was not even a disputed theory. They record that

> *"Dewey (1938:402) compares conceptual frameworks to maps. Maps are problem-solving tools. They help navigation through experience or the experiential world. They also represent and abstract from reality. When accurate, maps enable navigation within reality".*

(Shields and Tajalli 2006:316)

Dewey (1938) discussed the conceptual framework concept, at some length, over several of his preceding pages. There was no *theoretical perspective* term found in this 1938 Dewey work. Shields and Tajalli's warning about maps, *"When accurate"*, also reminds us that not all problem-solving tools actually solve the problem that they were designed to solve. It is a wry, cautioning observation that not all of these tools may actually deliver the claims made for them; *caveat emptor* once again.

The impact that assessment has on students has been encapsulated in this oft quoted, and it seems, often misquoted, observation:

> *"Students can, with difficulty, escape from the effects of poor teaching, they cannot (by definition if they want to graduate) escape the effects of poor assessment".*

(Boud 1995b:35)

Even the Dearing report (1997: chapter 9 paragraph 37) seemed to misquote it. It did however cite a different Boud paper. This is a powerful concept. It does seem likely that its author has

expressed it using a slightly different vocabulary, on different occasions, to different audiences. Its influence, whether the quote is as he wrote it, or paraphrased, is not lost. It is at the root of this study. Implicit in his assertion is that if student GSA is used then practitioners must ensure that it is used fairly.

The complexity is in determining what is fair. Simonite for example explained: *"Questions of fairness arise because students' degree awards should reflect individual achievement rather than the assessment methods experienced in their programme"* (Simonite 2001:262). Earlier in her thesis, she gave a rather pointed reminder that, *"assessment is intended to measure student achievement, not institutional resources"* (Simonite 2001:49). It seemed to be an argument thread worth developing. It is a shame that she did not have space to develop it, and in addition, it was not her main research interest at that time.

Fairness is a subjective, and context specific concept. Treatment of one person through a particular method may be fair, while the same treatment applied to someone else in different circumstances may not be fair. It seems likely that, for example, fairness for one person having the freedom to act a certain way will often conflict with fairness for another to be free from the effect of that act. Especially with respect to GSA, there seems to be no universal definition of fairness in an education context. The '*Standards'*, for example, note that even although

> *"Concern for fairness in testing is pervasive … . Absolute fairness to every examinee is impossible to attain … tests have imperfect reliability … validity in any particular context is a matter of degree".*

(Baker et al. 1999:73)

They also hypothesized, in their section on *Varying Views of Fairness* that the word has *"no single technical meaning"* (Baker et al. 1999:74). It has several different contexts. These, they noted, will often conflict. Additionally, the *Standards for Educational and Psychological Testing* (Baker et al. 1999) view on fairness is reviewed further in section 4.9 of this thesis.

An awareness of named theoretical perspectives enables researchers to evaluate the implications of their own research as well as that of others. As May (2001) asserted, it allows them to answer the question: *To what extent is any particular perspective applicable to my own research?* A named theoretical or philosophical perspective in a study report will allow its readers an insight into the beliefs, values and biases of its author. It explains the background to how the world looked to them when they formulated the study. In addition, it explains how they made sense of events while

conducting it.

While conceptualizing the research problem, researchers bring to bear not only their values as a researcher but also their personal philosophy of life. None can isolate their work completely from their life experiences. Neither can they work in isolation. This is equally true of this researcher. It has been noted that *"The research enterprise cannot be separated from the researcher"* (Kemler and Thomson 2006:60). The researchers even insisted, immediately following this, *"it is imperative to put the personal on the agenda through doctoral study".* It has also been noted that

> *"Scientists cannot divorce themselves from the cultural, social and political context of their work. What scientists can do is make their assumptions about the world explicit and strive to conduct their research as rigorously as possible".*

(Bowling 2002)

This text is also cited by Dee who added, *"researchers cannot separate themselves from the personal milieu from which they hail"* (Dee 2008:96). A summary of this researcher's work, academic study and research background is in section 2.7.1

Researchers' methodological procedures are influenced by whom they are. They should be constantly aware that both their *weltanschauung* and their life experience influence their thinking. They each influence the other. These personal views and experiences may also obfuscate any issues that might arise from some aspects of their studies. For example, this could apply to their methodology or ethics in how they deal with the problems in their own research. (Craib 1992)

Although researchers should strive for value freedom, it is naïve to assume that this is always fully achievable (Bowling 2002). In addition, as Taylor noted with regard to trying to avoid bias, when he was discussing political neutrality:

> *"There is nothing to stop us making the greatest attempts to avoid bias and achieve objectivity. Of course, it is hard, almost impossible, and precisely because our values are also at stake. But it helps, rather than hinders, the cause to be aware of this".*

(Taylor 1985:90)

In this section, this researcher's conceptual framework will be discussed against the concept of theoretical perspective. It will include a position statement on the meaning of a first academic degree. The term, preferred and used throughout this thesis, is *first-degree*. It should not just be assumed to be either a Bachelors degree or one of three years, full-time study, or appropriate part-time duration. It could be an integrated Masters over four or more years of study or a specialist or

professional degree (e.g. Midwifery) studied intensively over a much shorter duration.

Section 2.1 is entitled Principal stakeholders. Section 2.2 is on the subject of theoretical perspective and value neutrality. It further discusses why this present study cannot have a theoretical perspective. Section 2.3 is this author's position statement on the meaning of the academic first-degree.

Section 2.4 refers to the present study definition of the terms programme and module. Section 2.5 defines the present study definitions of ability, assessment and evaluation as they apply to this conceptual framework. Section 2.6 concerns some problems with summative assessment. Section 2.7 is this author's response to the question of why the quantitative affects of the GSA marking is worthy of study. It is a personal perspective, and it includes a work, academic study and research profile in section 2.7.1, and issues of assessment and degree classification fairness. The final section, 2.8, is the conceptual framework chapter summary.

2.1 Principal stakeholders

Principal stakeholders of GSA include not only students and higher education institutions, but also all of society. (A more comprehensive list is in Appendix 1.) It also seems unlikely that, as individuals, the stakeholders will all belong to only one category. It is quite likely that some of them may have more than one role as stakeholder. For example, a parent may also be a university teacher, a member of the Local Education Authority (LEA) and a local government official concurrently. In addition, they could also be a part-time PG student.

In the UK, the range of principal stakeholders includes anyone who pays, or has previously paid, income based tax in the UK. They have a stake in the funding of degree programmes, now and in the future. They may also be concerned to know whether undergraduates have the opportunity to both learn, and to develop. More specifically, have they learned and developed in ways that adequately equip them for their own future, which is also for the benefit of society. The future of graduates includes working and being taxpayers themselves. Some of this revenue may in turn, be used to support both the next and previous generations of taxpayers. The former are those who have yet to graduate. The latter are those who are retired and perhaps no longer contribute so directly to the national exchequer. Although the former category seems likely to be increasingly under pressure due to rising tuition fees. In addition, principal, or critical stakeholders may have concerns in relation to methods of measuring, and of students learning from, GSA modules.

Methods should be both adequate and fair. They also need to be understandable when the marks derived from them are included in graduate transcripts or Higher Education Achievement Reports (HEARs).

Summative assessments are judgements about a student's learning and attainment for the purposes of study progression or certification, made by markers who are also principal stakeholders themselves. The synthesis of these is the degree classification awarded to the graduate. Markers are also employees of principal stakeholders. The principal stakeholders of the GSA versus IISA problem include the whole of society. This is borne out by the 1985 revision of the APA *'Standards'*, which described these stakeholders slightly differently:

> *"Educational and psychological testing involves and significantly affects individuals, institutions, and society as a whole. The individuals affected include students, parents, teachers, educational administrators, job applicants, employees, patients, supervisors, executives, and evaluators. The institutions affected include schools colleges, businesses, industry, and government agencies. Society, in turn, benefits when the achievement of individual and institutional goals contributes to the general good".*

(Novick 1985:1)

2.2 Theoretical perspective and value neutrality

Theoretical perspective is specific to each individual researcher. Creswell (2003) explained it with his rainbow metaphor. He commented that it helps researchers, *"to visualise how a theory operates"* (Creswell 2003:120). Other researchers have explained that theoretical perspective was *"not merely a theory or a paradigm or a set of hypotheses but all three"* (Alford and Friedland 1985:389). Although their work was on a topic other than education, the Alford and Friedland explanation seems appropriate to this one. A researcher's theoretical perspective is grounded in their view of the world, how and what they think about how it works, and their outlook on it, at the time of their study. For example, they may favour a quantitative or a qualitative approach irrespective of the research topic. While their research approach may develop and change over time, it seems likely that it will remain stable between different research threads run concurrently.

Value neutrality, on the other hand, is the researcher's ability to keep an emotional detachment in their research. The extent of it may be different for each new research topic. It could either be a problem, or it could motivate the research if the topic was to parallel an important personal experience for the researcher, as noted in Strauss and Corbin (1990). See their quotation in section 2.7.1. Some researchers may think that they can be emotionally detached from the research topic. This seems unlikely, see the Bowling (2002) quotation in the chapter introduction.

In addition, theoretical perspective also encompasses the concept of value neutrality, rather than vice versa. Theoretical perspective is the label given to the researcher's general cognitive framework. It connects, however loosely, to a named, published theory. It is the researcher's position on the place of their research in relation to both the named theory and the rest of the world. This is similar to the Creswell (2003:120) rainbow metaphor, mentioned at the start of this section. This present study sits within summative assessment in the overlap of higher education, GSA and independent individual summative assessment. This is illustrated in the Venn diagram in Figure 1.

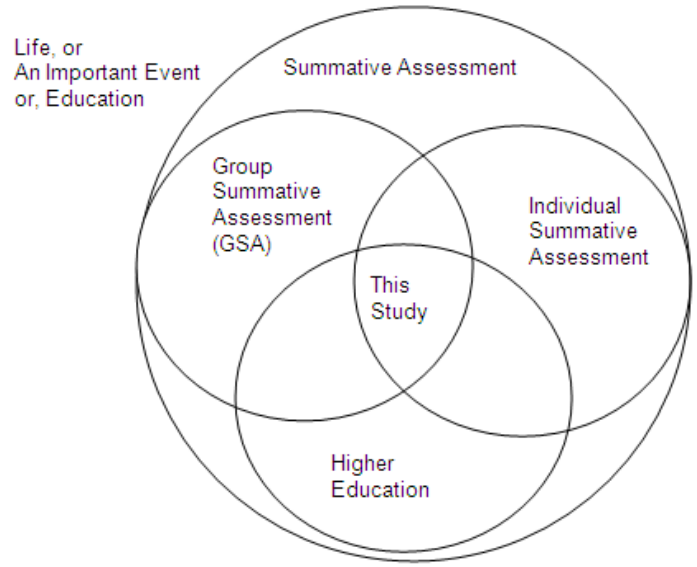


Figure 1: Focus of this study

A theoretical perspective encompasses researchers assumptions, untested ideas, hypotheses and theories. It is not specific to the researcher's current topic. It is transferable to other topics studied by the same researcher. It may develop, or change, over time. It also involves discussion, expectation and prediction regarding the research. It affects researcher's interpretations of their findings. It also affects their predictions of the impact that the independent variables may have on their study results. (See Figure 9, and Forrest-Presley et al. (1985)).

Theoretical perspective is also known as *"relativity of perspective"* (Phillips and Burbules 2000:47). Another term used to describe it is *"theoretical rationale"* (Creswell 2003:120). It is a personal ideological, even idiosyncratic, view of the world. It includes all theoretical approaches. For example, it includes quantitative and qualitative, Christian, post-positivist, feminist and Bayesian concepts, among others. There are also sub-types and combinations of these. They include belief and faith systems. They may covary in any combination to produce the researcher's unique

theoretical perspective.

Where a suitable named theory exists, theoretical perspective guides researchers to formulate the research problem. They can form the research questions, design their study, and carry out their research around the framework it provides. For them to have formulated or even adapted it, requires an extant named theory (Creswell 2003). As mentioned elsewhere in this thesis, there are no named theories specifically concerning the quantitative impact of GSA marking. Because of this, this researcher cannot have a *theoretical* perspective. On the other hand, this researcher can and does have an approach, or a conceptual framework. A similar logic applies to the concept of a theoretical *framework*. Hence, this chapter, section title and discussion are under the umbrella of *conceptual framework*.

Within the researcher's conceptual framework is the concept of value neutrality. Attaining and maintaining it should be a constant research aim. If researchers allow their feelings to influence the direction and analysis of their research, there will be no value neutrality. It has also been suggested by Phillips and Burbules (2000) that value neutrality is itself a value.

The researcher's background has an influence on the kind of person they are. This affects the form their research takes. The extent that any researcher theoretical perspective, conceptual framework or study method might be value neutral is unknowable. In addition, readers may not know, or even wish to know, details of the researcher background. This does not mean that the concept of value neutrality in the conceptual framework can be ignored. Researcher attributes and beliefs, all of which may have an impact on their research, need recognition, description and where necessary debate, in their published report.

2.3 What is a first-degree?

As mentioned in the chapter introduction, the preferred term for a first degree in this thesis is *first-degree* rather than for example Bachelors, BSc, or MSci. This is because the graduate's first degree may not always be a Bachelors degree. It may be an integrated Masters degree. It applies to qualifications awarded by all accredited degree awarding institutions. These are usually universities. This position statement includes the concept of GSA. It also applies to other participants in, and consumers of, higher education, i.e. its stakeholders (see section 2.1 above). A discussion on what a first-degree is a record of, and its interpretation and use by stakeholders is included. (Lee 2004).

A first-degree is firstly of course, a programme of study. On graduation, it is evidence of a qualification. It is a parchment, warrant or token. It is a passport into graduate professional employment, or further postgraduate or professional study. For the graduate, it is their first portable academic qualification. From the majority of HEIs, although not all of them, it has international recognition.

Some professional employment options may require particular degrees as an entry prerequisite. For others, almost any subject discipline will do. A first-degree in any subject may be qualification enough for a whole raft of professional opportunities. For example, a first-degree in Geography may be an entry qualification for careers as diverse as Finance, Information Technology, Distribution and Logistics, Teaching, Earth Sciences, or the armed forces and emergency services. In addition, a first-degree is usually a prerequisite for postgraduate study, and could be entry in an entirely different discipline of study.

A degree classification on the other hand, is an award made to students on their graduation. It is a proxy for their aggregated module summative assessment marks. However imprecise, it is a single measure. It is a permanent record. It represents the level of the student's academic attainment in that assessment at that time, as judged by their collective assessors. Individual students receive it. The marks are a conflation from those from the study modules in their degree programme. It is **not** an award made to student groups.

While studying for their first-degree, there may also be additional distractions from their studies that students may need to address. These could include anything from living independently from parental supervision for the first time, up to and including working part- or full-time to support their studies. Some may also have to contend with providing financial, physical or emotional support for dependents, or any or all of these, during their studies. A degree classification also reflects how they coped with these additional distractions and responsibilities; however, this is unmeasured and goes unrecorded.

Graduating after a first-degree is also a public acknowledgement by the HEI that the candidate has fulfilled all their academic engagements, including their assessments. If they are postgraduate students then it also means that they have engaged in research and/or advanced study to a

satisfactory standard. In addition, it means that graduates have no outstanding financial obligations to the HEI. They have no fines due on library loans for example, or overdue tuition fees, (Terms 2009).

Collectively, students must have a whole gamut of reasons for studying for their first-degree. For most of them, these will presumably include graduating. They may also include submitting to pressure (or encouragement) from peers, school or family. They could also include thwarting or confounding these pressures. A change in professional, domestic and/or financial circumstances may provide an opportunity and a prompt for academic study. Alternatively, it may just be in order to delay decisions or actions relating to other life choices. Degree study may also be a first generation opportunity. This is a variation on family pressure. Entry to a hoped-for professional career may be possible by attaining it. Some students may turn to academic study in order to avail themselves of the social, networking or travel, or even for the sporting, sports coaching, and training opportunities, it offers. It may also be seen by some simply as a personal challenge. (These and other reasons are included in Appendix 2.) All these individual reasons will covary to render each student's reason unique. Newstead (2002) found some of them in his research. He did not include acceding to, or confounding peer family or other pressures as a reason. His three categories were Stop Gap, Means to an End, and Personal Development (see Table 2).

Table 2: Reasons for degree study

Stop Gap (10%)	Avoiding work. Allowing time to decide on career. Social life and enjoyment
Means to an End (66%)	Enhancing career prospects. Getting good qualifications. Improving standard of living
Personal Development (24%)	Reaching personal potential. Gaining knowledge for its own sake. Furthering academic interest
(Newstead 2002:70)	

Most HEIs usually award four classifications of honours degrees. For example, at Durham University these are First or Honours Class, Upper Second Class, Lower Second Class, and Third Class. A description of the attributes expected of a Durham University Honours Degree graduate is on the undergraduate qualifications descriptors web page, (Descriptors 2008). To graduate, students will have developed an understanding of a complex body of knowledge. They will also have demonstrated a systematic understanding of their field of study, consisting of

> *"A conceptual understanding …*
> *An awareness of current disciplinary boundaries and an appreciation of the uncertainty limits and contested nature of knowledge within their programme of study.*
> *The ability to direct and manage their own learning effectively across a range of topics,*
> *Making use of scholarly reviews and primary sources (e.g. refereed research*

articles and/or original materials appropriate to the discipline).
The ability to undertake, with supervision, independent investigation of a defined topic within their programme of study and to report the findings effectively".
(Descriptors 2008)

Typically, the authors assert, a holder will also be able to *1) apply the methods and techniques that they have learned and 2) communicate effectively at all levels.* They will also have academic qualities such as flexibility and discrimination, initiative and personal responsibility, and decision-making and learning ability. There is also a footnote on the Descriptors (2008) website explaining that these qualification descriptors are generic. They apply to all subject areas at the relevant level across the University. A caveat explains that departments supplement these qualification descriptors with their own subject-specific descriptors. These take into consideration relevant benchmarks and other requirements associated with the discipline. This could include those needed for entry-level professional registration, e.g. the British Computer Society (BCS), see section 2.7.3. (Other examples of attributes of a Durham Honours Degree Graduate's systematic understanding are in Appendix 3.)

2.3.1 Andragogy - not pedagogy

In the literature on university education issues, the term most often used to refer to the science of teaching, is pedagogy. This is fine for students taught by pedagogical methods. That is, as though they were children. Otherwise, the preferred term should surely be andragogical. This point is raised here first rather than in the literature review or study design chapters because it has only an indirect bearing on the study topic. It was however sufficiently distracting in the educational literature on student group working for it to be included in this conceptual framework chapter.

Smith (2008a) wrote that Eduard Lindeman described his *" orientation as 'andragogical … - which appears to be the first English-language use of the term"* (Smith 2008a), and McKenna noted that Knowles *"developed the theory of andragogy, the art and science of teaching adults"* (McKenna 1995:31). These illustrate their authors' disquiet with the use of the word pedagogy when referring to university students, many of whom may bring a richness to their activities that leaves the university a better place for their having studied there. Changing one word for another in referring to their teaching is neither mere semantics nor fashion. It will encourage a change of the concept of a student, in the minds of students, university staff and other stakeholders.

In evolutionary terms, the difference between andragogy and pedagogy makes good sense. Adults, with their greater experience of the world would be able to promulgate warnings and advice

to other adults. The question is where do university students fit in this? Transactional analysis (Rogers 2001) examines the interactions between the child, parent and adult or of people adopting those roles. Even without these considerations, referring to university students teaching as pedagogical, rather than andragogical, is unjustified. Cognitively at least, university students are not children, irrespective of their chronological, emotional, or other age.

2.4 Study definition of the terms programme and module

In this study, the word *programme* will refer to any one of the suite of first-degrees offered by a degree awarding institution. Upon successful completion of one of which, a degree classification is awarded. One *programme*, studied successfully, leads to an award of one degree.

Several *modules* successfully studied during each year of the programme allow the candidate to progress their studies and ultimately to graduate and be awarded a degree classification. *Course, module* and *programme* are often interchangeable in the literature from different researchers and HEIs. A *course* can mean either a module or a programme, depending on the educational institution. *Coursework assessment* will however designate all non-traditional assessments, for example, an assessment that is not a timed, written examination, or a viva. (For an example of where *coursework* lies in the assessment hierarchy, see the programme assessment structure of the C data set in Figure 7.) Several *modules* make up a degree *programme*. The number of contributing modules in a programme depends on the HEI and the type of first-degree, e.g. Bsc. or MSci. It also depends on the modules weighting. For example, a first-degree could be a Bachelors degree of 36 modules weighted at 10 credits and studied over three years or pro-rater for part-time study. It could also be of fewer, and differently weighted modules, both variations would total 360 credits.

2.5 Study definitions of ability, assessment and evaluation

Throughout this thesis, it is assumed that *ability* is a proxy for the cognitive attributes that a student brings to an assessment. There are many ways of estimating this. They include both quantitative and qualitative methods. Using students mean A-level or IQ test scores would provide a quantitative measure of their prior *ability*. On the other hand, an assessment of the quality of the students ability to engage with the topic could be either quantitative or qualitative depending on the marking criteria. For example, it could be qualitative if it was assessed during seminar or tutorial teaching sessions as a high, medium or low category.

Assessment and evaluation can apply to either the module efficacy or the students' attainments in them. Whilst it seems to depend on the author, there seems to be a modern preference for assessment to refer to student attainment. Such is the intention throughout this thesis. The exception is when the work of other authors is presented, for example, in section 4.5.1, Scriven (1983) and Newton (2007). Earlier work by other researchers, e.g. Scriven, seems to favour assessment to label judgement of programmes. Evaluation commonly described measuring student attainment. At this time (1983), GSA would have been far less common.

2.6 Some problems with summative assessment

There seems to be a consensus among educationalists that summative assessment is *assessment of learning* or *assessment of attainment*. At least one author has noted the dual role of summative assessment. It has been observed that assessment has to *"encompass formative assessment for learning and summative for certification"* (Boud 2000:8). This duality does not aid clarity because the two are not mutually exclusive. Summative assessment also has a formative role. There are several additional definitions of summative assessment in section 4.4, and see for example sections 4.2, 4.5.1 and 4.10. Applying both roles to the same piece of work could mean that they were not always compatible. For example, some students might want to rehearse or test an assessment question response in a formative assessment that is not a threat to their module marks, or has any impact on their degree classification. In some group project modules, some students may also be competitive rather than collaborative in character. Institutional pressures on resources may also mean that the summative role of assessment takes precedence.

This section is a summary of general problems of summative assessment as a whole, rather than of those of GSA within it. It is here rather than in the literature review section; because part of the researcher conceptual framework is that there is no ideal summative assessment method. There are serious difficulties with most summative assessment methods. The following is from the literature on difficulties with examinations, essays, coursework, and GSA projects.

Of examinations, it has been noted, tautologically, that *"unseen written exams favour candidates who happen to be skilled at preparing for and sitting such exams"* (Race 1998:151). Another author was more pragmatic. They noted that *"Examination-orientation does less harm to those who are good at examinations; the learning process is less likely to be suffused with anxiety and fraught with frustration."* They added, *"Others may not be so lucky"* (Dore 1997:xxiii). In essay

marking Hartog and Rhodes (1935) seem to have been the first, in Britain at least, though not in North America, to highlight difficulties with reliability.

Coursework includes both IISA and GSA items. They are both susceptible to plagiarism and other forms of cheating. In addition, GSA can suffer from non-contribution from group members (Smithers 2006). This is also termed the free rider, or social loafer problem. This is discussed further in the next chapter, in section 3.9.7.

2.6.1 Slide-effect, contrast-effect and gender bias

The slide-effect was a problem noted by Griffiths (2002). He asserted that it concerned essay marking. In a newspaper article he noted, *"essays marked earlier in the sequence obtained higher marks than those marked nearer the end"* (Griffiths 2002). This could also apply to all other assessment methods, not just to essays. There was however, no peer-reviewed publication found that used the term. This was despite personal communication with Griffiths (2008), and with Newstead (2008), the author he subsequently suggested. (Also, see Meadows and Billington (2005:21), part of the text of which is in Appendix 4.)

Newstead confirmed that he drew on Farrell and Gilbert (1960) for his article. He also confirmed, *"the **variance** of the marks for students later in the alphabet (with names beginning L to Z) was higher than that for candidates earlier in the alphabet (with names beginning with the letters A to K)"* (Newstead 2002: emphasis added). He noted, *"The authors attribute this to the growing confidence of the markers as they become more familiar with the types of answers produced, but they acknowledge that it may be due to other factors such as fatigue"* (Newstead 2008). He also confirmed that he did not know of any other work in that area, and was *"unfamiliar with the term 'slide-effect'"* (Farrell and Gilbert 1960; Newstead 2002).

As Newstead indicated, the Farrell and Gilbert finding was subtly different to the Griffiths interpretation. They had reported that it was the *variance* of the marks that increased with the number of scripts marked, rather than the marks themselves.

Anonymous marking may also have an impact on any slide-effect. The order of marking will be unknown. If the slide-effect does exist then it only moves the problem. It does not reduce it. It simply means that any slide-effect will be random throughout and within the students marked work. This only exacerbates the injustice that might be caused by the existence of any slide-effect. In

addition, the marking may be less anonymous than it might appear to be. The module convenor is the academic team leader. They, or other members of the teaching team, may be aware of the students essay topics, despite so-called anonymous marking. From this author's personal undergraduate, post-graduate and HEI teaching experience, negotiations between the student and their teacher on a suitable essay or project topic are often a module requirement of the assessment item. This means that the marker would be able, consciously or unconsciously, to match the topic, and therefore the work being assessed, to a particular student. The summatively assessed work may also have been the topic of a previous formative assessment exercise, which could also lead to the student being identifiable. A formative mark given to the student together with feedback could or should lead to improvement in the summative version of the same topic. It could also reduce or negate the anonymity aspect of the assessment marking. It is a separate issue, not pursued here.

Slide-effect would be a useful label for such phenomena, but there seems little support for it in the literature. The finding does seem quite plausible, similar to the contrast-effect (next paragraph). It should be the subject of future research.

The occurrence of a *contrast-effect* is another difficulty with summative assessment. Ebel (1961) seems to have been first to note this. It is not quite the same as the slide-effect. It refers to the effect on the mark due to the quality of the work marked immediately before it. For example, good work may receive an even better mark when it is marked following poor work, and vice versa.

Gender bias in assessment marking has been studied using A-Level examination data (Baird 1998). The author was able to rule it out in non-anonymous marking. If this finding is generalizable to higher education, then this potential problem may not be a problem after all. Research on the issue is underreported, possibly from the file-drawer effect (see section 2.7.2).

2.7 Why study the quantitative impact of GSA marking?

Trafford and Lesham (2002) asked the generic version of this title question in their paper on preparing for predictable viva questions. They simply asked, *"Why did you choose this topic for your doctorate"* (Trafford and Leshem 2002:1). A response, at least for this study, is less easily articulated. Concern for the research problem (section 1.1), which began around 1997, has not yet abated. For this researcher, at that time unemployed and studying for an NVQ while reflecting on

the issues involved with degree-level study, there was something indefinably disquieting about the concept of GSA and more particularly its practice. The disquiet began whilst perusing the prospectuses of local universities.

Key words derived from the GSA acronym, as well as GSA itself, have been *'Googled'* intermittently ever since then. As it became available, an RSS feed was also set up to provide automatic advance warning via electronic mail of possibly relevant articles from the academic journal whose content mirrors this researchers interest most closely. These on-line searches have indicated that the issue surrounding the impact of GSA on overall marks has remained a non-problem, at least as far as more traditional, mainstream educational academic research is concerned. Therefore, the topic chose itself. The main reasons for choosing this study population are explained in section 6.1.

The following reflections are from Phillips and Burbules; Creswell; Coe, and Paxton. They help to describe in part this researcher's outlook, his passage through society generally, and personal background. These all have an impact on the rationale for this current study. Some authors, for example, have noted that *"beliefs can be false beliefs"* (Phillips and Burbules 2000:2). They explained that what might seem to be *enlightenment* can be false. Understanding can be misunderstanding. They added that *"A position that one fervently believes to be true - even to be obviously true - may in fact be false"* (Phillips and Burbules 2000:2, original emphasis).

The prompt for this study was this author's perception of GSA practice in high-stakes public qualifications. It further evolved from personal experience as a student on both a first-degree and taught postgraduate programme of study. The belief was, at that time, intuitive rather than evidence informed. It was that GSA was somehow treating students unfairly, and may be false in its assumptions. Students subjected to it were being treated differently to those being individually assessed therefore their treatment was unfair. A little while later, after experiencing GSA from the other side, i.e. teaching undergraduate GSA modules, this view strengthened. Curiosity was further aroused. This was a topic worthy of proper research, with appropriate intellectual rigour and resultant academic reliability.

Lejk et al. (1999) and Hoffman and Rogelberg (2001) seem to have been the only other researchers whose published work has included findings on the effects of GSA. This was not the

only or even the main, focus of their research. In addition, Knight (2004) and Simonite (2003a) have published separate studies that were parallel to this topic and with a slightly different context. Both authors are cited later.

This researcher believed that alternative hypotheses, theories and debate would only develop from a research thread. It was missing. Someone had to start it. No one else seemed to be concerned. It has been noted that *"scientific advance proceeds by the accumulation of knowledge, not by results considered in isolation"* (Coe 2004). Another researcher has noted that *"Theories develop when researchers test a prediction many times"* (Creswell 2003:120). These observations also encouraged this study.

On a more personal note, Tom Paxton said it best. He recorded a song that included the line: *"It's a long and a dusty road; it's a hot and a heavy load".* He included *"I've been around this land, just a-doin' the best I can, Tryin' to find what I was meant to do"* (Paxton 1963), these sentiments mirror this researcher's thoughts on this study.

2.7.1 Researcher work, academic study and research background

When researcher background is included in a report, it allows the reader an additional insight into the motivation for both the research, and for the methodology of the research. Hence, it aids understanding of the text (Alkin 2004a). In this example, it was the researcher work history, as well as the group working and GSA experience during full-time undergraduate and early, taught PG study, which prompted this research. A researcher's background may also indicate where their potential for bias arises and how it may manifest itself. Alkin (2004a) explained why consumers of research literature also need to have an appreciation of their authors' journeys through society:

> *"To know someone's roots is to better understand the lens they use to view the world and the factors that might lead to initial views changing or being reaffirmed".*
> (Alkin 2004a:373)

This author's research interest is in the impact that GSA marking may have on an individual undergraduate's overall mark. It came about from the firm belief that a first-degree classification is personal to the individual graduate. It is also a high-stakes qualification (see section 8.2.3) and a very public one. It is an award for, and an acknowledgement and measure of, effort, ability and attainment, however imprecise and badly defined as mentioned earlier. No part of the individual's degree classification should directly depend on the efforts and abilities of one or more small groups of their peers. Students could not be sure they were receiving their peers best efforts and even if

they were, the results of those efforts may not be adequate. In addition, the recipients of any subsequent degree classification could not be sure whether the same award would have resulted from their own independent individual efforts and abilities. The question would always remain, was the award deserved or would they have received a higher or lower classification if they had been assessed individually. Since this researcher's first two degrees had strong GSA input, this latter point was rather important.

Concern for the topic was a necessary, but not sufficient prerequisite for this author to have begun to contemplate any degree by research. It seemed that the only way to find an answer to the research problem was by becoming a researcher and researching it! A recognised high stakes award of PhD at the end of the study was also important. This is because other stakeholders need to consider this study's findings seriously, and act on them. Critical stakeholders include educational academics and policy makers. The award could influence how the education community receives this thesis. The topic and doctoral study share an almost symbiotic relationship for this researcher. Without the concern for the topic, there would be no need for a doctorate. Equally, a doctoral degree requires a contribution to knowledge from the candidate. This thesis is such a contribution.

This researcher was new to research. Despite this, the presence or absence of a doctorate qualification will not affect this researcher's personal development. Retirement looms. What follows in the remainder of this section is an attempt to convey the intense interest in the topic. It contains elements of the researcher as researched, especially in terms of reactions to the GSA process and concept. At the time of the genesis of interest in the topic, the overriding priority was to secure a place on an undergraduate degree programme. A *certificate* in computing and an in-progress NVQ might not have been enough to satisfy the admissions tutor. As it turned out, they were.

Several authors have gone into print with their interpretation of the reasons why researchers are researchers, i.e. what motivated them to take the research path. In this Strauss and Corbin passage, the word *human* is as printed. It does not seem to be a typographical error on the word *humane*. Perhaps it should have been.

> *"Professional experience frequently leads to the judgment that some feature of the profession or its practice is less than effective, efficient, human, or equitable. … Some professionals return to study for their higher degrees because they are*

motivated by that reform ambition".
(Strauss and Corbin 1990:35)

This current study was not just motivated by professional experience. It was because of the combined experience from both academic study and professional career. The section from which the Strauss and Corbin passage is taken is headed *"Personal and professional experience"*. The authors continued the paragraph by observing, *"The research problems that they choose are grounded in that motivation"* (Strauss and Corbin 1990:35).

As a university student, all of this researcher's GSA groups were self-selected. Even so, GSA experience on both the BSc. and the MSc. programmes was both challenging, and confusing. Almost certainly, this was in ways that were unintended by the module designers. As perceived at the time, and reinforced after more than a decade of reflection since, the academic part of the group working and GSA experience of both programmes was, in addition, rather indifferent.

While teaching an undergraduate module that included a GSA item, students were allocated to their group according to this researcher's pseudo random allocation algorithm. The method was explained to the students during class time. The intention was to revise the group allocation method where necessary, following discussion and feedback on the proposed group allocation method. There were no comments from any student in any of the seminar groups. This salutary experience provided additional evidence for the structure of the research problem that subsequently led to this study.

Strauss and Corbin (1990) were correct though, at least in the case of this study. Experience did lead to concerns regarding the practice being less than *effective, efficient, human, or equitable.* The *reform ambition* certainly motivated this study. It was oddly reassuring that Strauss and Corbin (1990) noted *"Choosing a research problem through the professional or personal experience route may seem more hazardous than through the suggested or literature routes. This is not necessarily true"* (Strauss and Corbin 1990:35).

Lofland and Lofland (1995) expressed a similar sentiment in their guide to qualitative observation and analysis. The noted that *"Many social science projects have had their genesis in current biography"* (Lofland and Lofland 1995:1). Later in the same book, although offering no evidence to support it, they suggested that, *"much of the best work … is probably grounded in the remote*

and/or current biographies of its creators". They elaborated on this at some length.

> *"It is often said among sociologists that, as sociologists, we "make problematic" in our research matters that are problematic in our lives. Within the proviso that the connection between self and study may be a subtle and sophisticated one, not at all apparent to an outside observer, we would argue that there is considerable truth in this assertion. In fact, much of the best work in sociology and other social sciences - within the fieldwork tradition as well as within other research traditions - is probably grounded in the remote and/or current biographies of its creators. That such linkages are not always, perhaps not even usually, publicly acknowledged is understandable; the norms of scholarship do not require that researchers bare their soles, only their procedures".*

(Lofland and Lofland 1995:13)

Alkin edited, wrote and co-authored several chapters of the book, *Evaluation Roots*. One of his chapters began with the hook *"In the immortal words of a noted seaman philosopher, "I y'am what I y'am".* He referenced this only with the word Popeye (Alkin 2004c:291). He followed it by a more academic reminder that *"We are all modelled by our experiences"*, adding that it does not matter *"whether they be educational (formal or informal), events, or interactions with people"* (Alkin 2004c:291). It is interesting to see that in this phrase he is also reminding us here that for him there is a distinction between formal and informal education. As part of the same paragraph, he explained that these words originally included "reflecting on what had led to *changes* in my views on evaluation over a 25-year period" (Alkin 2004c:293, original emphasis).

A pilot study was instigated as a subsequent MA dissertation project (Almond 2006). It used existing marks data from a single cohort of undergraduate Science Faculty students at a university in the northeast of England. Figure 2 is a dual line (see section 7.4) regression chart of the IISA and GSA marks data.

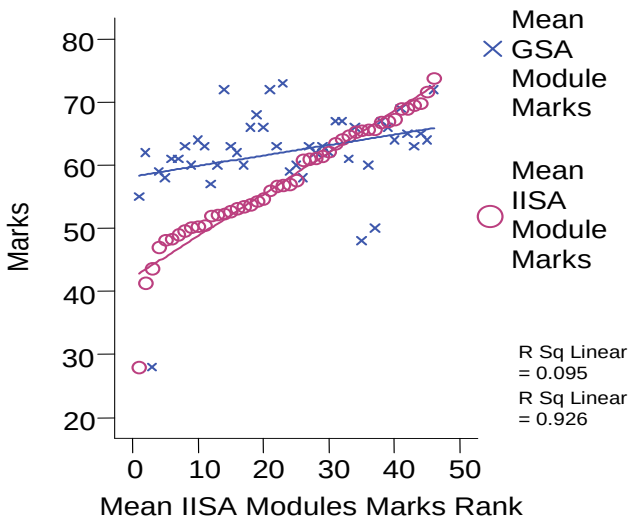


Figure 2: Scatter plot of group project (GSA) and mean other (IISA) module marks

This pilot study predicted that the least individually able students would have better group

assessment marks than independent individual marks. The more able students would have lower group assessment marks. Results from this pilot study were published in Almond (2009). From this, the ethics and morality of using GSA in university degrees are doubtful unless there are additional supporting considerations due to the nature of the study programme discipline, assessment topic or context.

2.7.2 It is an underreported area of educational research

Research into the effect of GSA marking on students overall marks is an underreported area of educational research. On the other hand, as previously noted, the research literature on GSA abounds with reports of methods of allocating students to groups. The literature also includes methods for deriving individual marks. They aim to ensure that methods of allocating individual marks are appropriate, transparent, reliable and repeatable. Above all, they are to try to ensure that the marks are fair, accurate and defensible. (Also, see Bibliography of methods of deriving individual marks list in Appendix 5).

Assessment that does not use GSA must assess students' *independent individual* effort and attainment. The methods used are, in some cases at least, centuries old. They may include timed written examination, viva voce, essay and dissertation. Contemporary individual assessment also includes individual coursework assessment. It could be IISA coursework essays or research projects. IISA is an accepted method. This seems to be despite, rather than because of any research evidence. (Also, see section 4.6).

The lack of published findings on the topic may be indicative of a lack of interest among educational assessment researchers and academics. Alternatively, however unlikely this may seem, it may indicate a lack of understanding of the GSA concept. It seems more likely however, that the lack of published findings is due to publication bias, the so-called *file-drawer* problem (see next paragraph). This occurs where researchers have not found anything that academic publishers deem to be worthy of publication. This could happen if researchers found that, in their sample, GSA had no effect on students' overall marks. If this was their conclusion from their research, then they might have difficulty in finding a publisher for their paper (Coe 2004). It has also been noted that *"unappealing conclusions tend to be ignored by researchers"* (Hammersley 2003).

The file-drawer problem is so-called because a research paper that validates a null hypothesis, i.e. one that reports that the study concluded that there was no difference, will often just end up there,

filed in a researchers drawer, ignored. On the contrary, if GSA was to have as similar impact on a student's overall programme mark that IISA has, and hence on their degree classification; this would be a very important finding worthy of publication. It has been pointed out that *"Such findings may be as important in research as any positive ones"* (LEMMA 2009: Module 2, section C.2.3.3).

The extent of any correlation between students module marks from individual work, and their mark derived from groupwork, whatever the method used to calculate this, was, with the exception of the authors previously noted, largely unknown. This study addressed this lacuna.

2.7.3 GSA fairness

One of the secondary reasons for the study was to explore the extent to which GSA is fair. As Simonite observed,

> *"no student should be disadvantaged by choosing one set of modules rather than another. Systematic differences in the outcomes of assessment by different methods raise questions of fairness to students who aim for the same award, but follow programmes that differ in terms of the assessment methods used to measure performance".*

(Simonite 2003a:460)

Is GSA fair to every undergraduate? Can it at least be as fair a method as we assume IISA to be? These basic questions prompted this study. A further question that arose from data analysis during the study is, is it fair to students of all abilities?

At least one author has likened group assessment to a game (Pitt 2000). He asserted that the rules of the game advantage some students and disadvantage others. In addition, he observed that,

> *"While the allocation of marks is a motivator, factors such as* ***`teamwork'*** *and* ***`contribution to the group' are hard to define*** *and essentially impossible to assess fairly".*

(Pitt 2000:239-240, emphasis added)

The 1999 revision of the *Standards* (Baker et al. 1999) specifically included a chapter on *fairness in testing and test use*. The *Standards* four distinct views of fairness are summarised from Baker et al. (1999:74-76):

> One: the lack of either item or test bias, i.e. no differential item functioning (DIF, which occurs when the probability of giving a certain response to an assessment question is different between different groups. It applies only when the groups have the same latent trait, i.e. range of skills and abilities. For example, these differences are commonly between genders or different ethnicity (Mahoney 2008)).
>
> Two: equitable treatment in the testing process, i.e. examinees having comparable opportunities to demonstrate their understanding on a construct and receiving fair and just treatment during the assessment process.

> Three: equality of outcomes across groups.
>
> Four: opportunity to learn, i.e. the extent individuals and groups have had adequate instruction or exposure to test content. (Camara and Lane 2006:37).

Camara and Lane also noted,

> *"The Standards dismiss the idea that equal outcomes across different groups are an acceptable criterion for evaluating tests and test use, but note that differential outcomes may trigger additional investigation to determine if internal characteristics of the test or its application could be biased against a group".*

(Camara and Lane 2006:37)

Section 4.9 explores the joint AERA, APA and NCME 'Standards' fairness chapter further.

The British Computer Society (BCS) is a professional institution. It accepts computer science graduates for entry-level membership. The BCS guidelines for a degree programme from a computer science school to be fully accredited included this clause.

> *"The project work may be part of a group project, but the technical report and assessment must clearly identify each individual's personal contribution and each student must clearly identify their contribution to the overall project, including analysis of how the group functioned and the role(s) played by themselves".*

(BCS 2007:14)

This shows the importance given to the individual contribution to the degree. There is also an implicit acceptance of the role of group work in the degrees awarded to its applicants for membership.

Downie (2001) noted that Gibbs (1995a),

> *"rightly highlights a key implication of including group work in university courses, stating* 'The main problem with group project work is that it is individuals who gain qualifications, not groups, and some way has to be found to allocate marks fairly to individuals'*"* .

(Gibbs 1995a:13; Downie 2001:7).

This is true whether it is a student group studying for an academic degree or a team in the real world of paid employment. The main difference may be that they use different marking schemes. In addition, it could apply to a group, or team, competing in some sort of physical or mental contest. The degree classification award is also external to any politics, power or prestige plays by members of the group. For example, prospective future presidents of the student union and members of parliament could practise their oration and communication skills on the group without affecting it.

The Gibbs quotation in the previous paragraph was from1995, yet, there are still no published reports of a universally accepted GSA method. This is despite the plethora of variations that

attempt to establish both transparent and fair methods. Could this be an indication of conflicting and maybe mutually exclusive aims? Perhaps they are simply irreconcilable.

Part of the rationale for GSA practice in HEIs is that it simulates a professional working environment and prepares students for a future career as employees (see section 3.5.2). This is by no means 100% defensible. In the real world, the employers, or rather their managers, allocate group membership, not staff. Employees are rarely able to choose their own group. Student academic GSA projects are very different to those of their real world workplace counterpart in their format, authenticity and working environment. This is important from the point of view of student GSA project realism in comparison to that of the real world workplace. For example, as mentioned earlier, it is extremely unlikely that very many company employees much less new graduate employees; would be free to select their own work group. The implication that in this context employment and study are sufficiently similar for GSA to prepare graduates for future employment is not entirely justified. (The rationale for GSA practice in the literature is reviewed in section 3.5.)

Exploring how employers allocate employee groups was not a part of this study. It has however been part of this researcher's work background over several decades. This was often as a manager in manufacturing industry, where team working was considered an important issue. The evidence was collected as a participant observer of the efforts employers made to allocate employees to project groups. Inevitably, they were not self-selected; senior managers selected them.

Evidence for this was from discussions with peers of the day. In addition, *team* rather than group, is the preferred term in the real world of paid employment (see section 3.1). Junior employees would have very little input on the *'team'* to which they are allocated. In addition, to swim against the group norm tide in a paid work environment could also be a career limiting decision for any employee. It would be very brave of a new graduate to make such a bold move. It seems likely that most of them would adhere to the '*not invented here'* syndrome, and resist proposing any novel action. It is also unlikely they would resist the tide of groupthink (Bion 1961; Janis 1972). It is much more likely that new graduate employees would comply with existing group norms than try to change them.

In contrast to this, the most common method found in this study for forming groups for student

projects, was self-selection. The rationale for this is articulated by several authors who observed that, *"in order for a group to function well, each member of the group must feel comfortable in that environment and be committed to the task in hand"* (MacBean et al. 2004).

The use of self-selected groups in student GSA may be more for risk management purposes than for andragogical reasons. It may simply be perceived as the method that entails the least business risk for the module providers. These risks include being criticised for using unfair group allocation methods. It may also be easier for the tutor to administer a GSA module if the groups select themselves. They may be more amenable to the process. This is supported by the MacBean et al. quotation above. It may also be a method of empowering students. Despite this, allowing student GSA groups to self-select will often negate any pretence at simulating real world work groups or teams. Employee teams, as mentioned earlier, are not generally self-selected. Where it might at least parallel the real world would be if the students negotiated with their teachers to formulate a method of allocating students to groups. Where students have only limited experience of groupworking, it seems likely that the choice of group allocation method would be the staff's by default.

Future studies on the fairness of GSA practice might include differences between students with different attributes, e.g. personality, gender, culture. They are not the focus of this study.

2.7.4 Stakeholder resource and research thread

Another strong reason for this present study was for the thesis to provide a GSA resource and research thread for stakeholders. This will encourage confidence that such methods may, when made transparent, be appropriate on a wide variety of degree programmes. Stakeholders will also be able to see that this research was examined and is replicable. It is also an attempt to provoke debate and reaction from critical stakeholders. It will add much needed additional evidence to the under-supported research thread. Results and conclusion from this study will lead to a better understanding of the effect of GSA and a greater confidence in the assessment method. They will also provide background evidence for longitudinal and meta-analytical studies. These are necessary because even though it seems that there is an imbalance of students IISA and GSA marks distribution; this present study seems to have been the only in-depth study of its kind.

Experience in groupwork projects and group learning will be greatly beneficial to undergraduates – after they have graduated (see section 3.5.1). It will help their preparation for the imminent and

competitive real world of paid employment in their chosen professional career. It provides practical experience, and even an understanding and/or an appreciation of group-dynamics. This is not sufficient to justify its use without evidence of the impact it may have on students overall marks.

2.7.5 Study for its own sake

This last part of the conceptual framework response to the why study the quantitative impact of GSA marking question relates to a rather less robust concept. It is also less easy to define. It is about study almost as a hobby, that is, for its own sake, as a reason for completing this thesis. The reasons are threefold, and although this section is quite short, is every bit as serious a reason as those preceding and succeeding it. Firstly, this study was of something interesting and meaningful to this author. It was for its own sake. The study required new skills. It certainly required new knowledge. Secondly, there is the wish to lay the ghost, the bête noir mentioned in the study introduction in chapter 1. The alternative could have been reflecting regretfully on *what if …* in the future. The third reason concerns a lifelong admiration for those who have gone before, i.e. those entitled, by the results of their studies, to use the title Dr. It has often seemed that most, though not all have a tolerance for the frailties of human nature that is an admirable aspiration. This is not sufficient reason for undertaking this project, but it was a necessary one.

2.8 Conceptual framework chapter summary

This chapter has described and discussed the researcher conceptual framework. Section 2.1 highlighted who the principal, or critical, stakeholders of GSA are. Section 2.2 was in relation to why this thesis has no theoretical perspective chapter and why it is entitled conceptual framework instead. It included an overview of the concept of value neutrality. Section 2.3, a position statement, was this author's response to the question of what is a first-degree. Section 2.4 explained the present study definitions of the terms course, module and programme. Section 2.5 defined how the terms ability and assessment, and evaluation, apply to this present study. Section 2.6 concerned the main problems associated with summative assessment. It included an explanation of the terms slide-effect, contrast-effect and gender bias. Section 2.7 responded to the question of why study the quantitative impact of GSA marking. It also included this researcher's personal work, academic study and research profile, in section 2.7.1. The section also included issues of GSA fairness. The apparent lack of concern in the research literature for the topic among academics was included in the discussion.

The next chapter is chapter three. It is the first of two consecutive literature review chapters. In it,

some of the broader, more general, educational background literature, in terms of GSA practice, is reviewed. The second and final literature review chapter will review the HE assessment literature.

Chapter 3. Review of general background literature

Hamerton, cited by Moore, observed sharply,

> *"Have you ever observed that we pay much more attention to a wise passage when it is quoted than when we read it in the original author?"*

(Hamerton 1887:142; Moore 1915: vi)

The previous chapter presented the conceptual framework of the study. It included an overview of who the stakeholders of the research were and how theoretical perspective relates to theoretical framework. The latter included an overview of value neutrality. This was followed by a personal reflection. It was a position statement on the concept of an academic first-degree. Additionally, it included speculation on what a first-degree is, or at least what it might mean, to different stakeholders. The study definitions of *course, module and programme* were also included in the previous chapter, as were those of *ability*, *assessment* and *evaluation*. The first of three remaining sections concerned the main problems of summative assessment. The second was the reason why this study was necessary. The third was a summary.

The concept of education and educational assessment, reaches back to Confucius, 551-479 BCE and Aristotle, 384-322 BC (Welton 1914; Huanyin 1993; Palmer 2001; Lambert 2003; Richardson 2004). There is however, a dearth of directly relevant contemporary published work on the quantitative impact of GSA marking on HE students' overall marks. This has been supported by other researchers who noted that *"published research related to collaborative learning in higher education is relatively scarce in this emerging field"* (Brown and McIlroy 2011: 688).

This is the first of two literature review chapters. Because of the '*relatively scarce published research',* both chapters are part of the wider background to the study rather than providing a basis for it. Literature on education in general, will be reviewed in this chapter. Group-dynamics in education literature will also be reviewed. (In the second literature review chapter, the literature on summative assessment in higher education will be reviewed.)

The topic of GSA is here in the *Background* Literature Review chapter rather than in the next section, Assessment in HE, because some of the literature and research findings apply equally to all students undertaking any formal learning programme, rather than just HE students. Some of the arguments about groupwork, and hence about GSA may be age independent. They may not be exclusive to just HE students studying for their first-degree. As well as university students, some

are generalizable to younger school pupils, and to sixth form and FE students. They also apply to both vocational and professional advancement courses. In addition, they may also apply to other areas of specialist education and training such as sports coaching, the armed forces or police and other emergency services. These other emergency services could include for example the voluntary Royal National Lifeboat Institution, and commercially run motor vehicle rescue organisations. In such organizations, the outcome measures seem likely to include both the overall success of the group and the performance of the individuals in it. GSA and IISA may even apply to formal and informal friendship, sporting and social groups. It would apply where the groups' members consider the stakes high enough. For example, where the group is competitive, internally or externally, whether or not they are amateur or professional.

Several authors have published work that included their rationale for, and findings and views on, group work and GSA in higher education (e.g. Candy et al. 1994; Gibbs 1995a; Morris and Hayes 1997; Downie 2001; Barfield 2003; Ackerman and Plummer 2004; Crebert 2007). In addition to the researchers cited in the opening paragraphs of this thesis, others have expressed caution for group work:

> *"Any perspective on group-work should emphasize that it is not appropriate for all learning occasions with all students".*

(Thorley and Gregory 1994:178)

They reminded us that different people have different ways of working and studying and added, *"We must make sure group-work is not used purely with the intention of saving staff time"* (Thorley and Gregory 1994:178-9). There is further discussion on this view in section 3.5.5.

Section 3.1 outlines and reviews the general background literature on the definition of group and team. In section 3.2, groupwork, group work, group and peer learning, and group assessment are reviewed. The literature on the Practise effect is reviewed in Section 3.3. Section 3.4 relates to The ubiquitous nature of group work and GSA in higher education. Section 3.5 reviews the Rationale for GSA practice literature. Section 3.6 concerns How student GSA groups are formed, including reviews of the findings from the literature on GSA heterogeneous versus homogeneous personality groups (section 3.6.3.3).

The limited literature on The impact of GSA on student marks is reviewed in Section3.7 and section 3.8 describes an unverified claim (Wollongong 2006). This concerns the relationship between low and high achievers group and individual marks, which tends to support some of the findings in this

study. Section 3.9 focuses on the literature on the Disadvantages of GSA.

Section 3.10 illustrates the views of student, staff and alumni on GSA while the rather limited degree classification research is the focus of section 3.11. The Southampton University *Jumpstart* group-working initiative is outlined and reviewed in section 3.12. Section 3.13 reviews the literature reporting biographical predictors of student GSA outcomes. Section 3.14 highlights the comparatively little used term *andragogy*. Finally, section 3.15 summarises the chapter summary.

3.1 Group and team: definition and difference

In the literature, the distinction between the terms *group* and *team* is often unclear, even interchangeable.

Mutch (1998) illustrated the difference between a group and a team rather elegantly. He discussed two models.

> *"a team-based one, aimed largely at preparing students for employment and a group-based one, aimed primarily at supporting the learning process".*
> (Mutch 1998:50)

His paper was on the place of student GSA experience in terms of both employability and learning in higher education. One of his sections was entitled simply '*Teams or Groups?'* He suggested that in situations where people work together as a group or team, there is a continuum of working together that have four characteristics. These were, he noted, permanent/shifting, formal/informal, authorized/hidden and assigned/self-selected. He also confirmed that the words *group* and *team* were often used interchangeably. This, he suggested reflected,

> *"both the blurring of boundaries along the continuum and a lack of clarity in usage, a lack of clarity which we can see reflected in higher education practice".*
> (Mutch 1998:51)

On the other hand, some authors have explained that all teams can be seen as groups, but not all groups can be teams. They used the term group *"for people who come together to share knowledge, for personal development or to learn from each other through discussion"* (Jaques and Salmon 2007:6). For them, a team was also a group who "*are engaged in a task or project geared towards an end product or decision"* (Jaques and Salmon 2007:6). There is a discussion of this in section 8.1.1. In an earlier book (2000), in a chapter on *Theories about Group Behaviour,* Jaques had also noted an additional group difficulty he had become aware of. This was the *"constant problems to do with the lack of motivation and commitment, alienation and even* 'dropout'*"* (Jaques 2000:9). He had even suggested, *"many of the issues may seem far-fetched to academic tutors"* (Jaques 2000:9).

Katzenbach and Smith (1993a) wrote on organizational teamworking. They warned that merely re-labelling *the collective* from group to team was not enough to ensure compliance with the concept. Teams, they observed, that perform adequately were

> *"not amorphous groups that we call teams because we think that the label is motivating and energising, … there is a basic discipline that makes teams work ... teams and good performance are inseparable, you cannot have one without the other".*

(Katzenbach and Smith 1993a:111-112)

They also reminded us that

> *"Groups do not become teams simply because that is what someone calls them ... A team's performance includes both individual results and … collective work products [that] two or more members must work on together. … a collective work product reflects the joint, real contribution of team members".*

(Katzenbach and Smith 1993a:111-112)

They further explained that in their view the essence of a team was the common commitment of its members. They argued that if the only collective activity the team engaged in were to meet occasionally just to make decisions, this would not be enough to sustain the team's performance. They defined a team as

> *"a small number of people with complementary skills, who are committed to a common purpose, set of performance goals and approach for which they hold themselves mutually accountable".*

(Katzenbach and Smith 1993b)

Their assertion that '*teams and good performance are inseparable'* would depend on their definition of good performance. Take a soccer match for example. Is good performance associated with entertaining, skilful soccer, or merely with winning? With the latter criterion, only one team in any competition could be perceived as giving a good performance. Katzenbach and Smith also noted that teams have a common purpose and hold themselves mutually accountable. These are necessary attributes of a team, but they are incomplete. Team, or student group, members also work together interactively. Their (1993b) definition omitted this.

In addition, their definition of *team* should have ended after common purpose. The remainder is simply embellishing the definition of the phrase *common purpose*. It is just paradoxical that they named, as a shining example of the use of teams, the innovative USA energy giant, the Enron Corporation. Despite its size, with $101 billion turnover in 2000, it famously went bankrupt in 2001. In addition, some of its directors were successfully prosecuted for fraud. The ex-chief executive Jeffrey Skilling, for example, originally received a 24-year jail sentence (Calkins 2004; Clark 2010).

The Enron Corporation could still be an example of a great team, just a great team of crooks!

It has also been noted by Sheard and Kakabadse (2004), after citing the Katzenbach and Smith definition of a team, that *"Essentially the same sentiments were also encapsulated by Babbington-Smith (1979)"* (Sheard and Kakabadse 2004:13). They also made the point that *"Groups and teams are not the same thing"* (Sheard and Kakabadse 2004:18). In addition, they asserted, *"the group has no need or opportunity to engage in collective work that requires joint effort"*.

Forsyth (2006) has listed eleven different authors definitions of the word *group*. Each definition was under a different *central feature*, e.g. categorization "*A group is* "two or more individuals … [who] perceive themselves to be members of the same social category"" (Forsyth 2006, original emphasis). His central features of a group also included communication; influence; interaction; interdependence; interrelation; psychological significance; shared identification; shared tasks and goals; structure and systems. He named and listed the characteristics of his basic types of groups and gave an example of each of them.

Oldfield (2005) also added to this debate. He asserted that, *"In much of the research and commercial literature the terms group and team appear to be used interchangeably … there is a significant difference between the two terms"* (Oldfield 2005:IV). He quoted another team of researchers as having asserted, *"purpose is fundamental to all groups, teams are specifically, deliberately, and invariably about results"* (Oldfield 2005:IV, 1-2).

The Sheard and Kakabadse (2004) common elements of teams, (cf. Much's (1998) continuum above,) were one, common purpose; two, interdependence; three, clarity of roles and contribution; four, satisfaction from mutual working; five, mutual and individual accountability; six, realisation of synergies; and seven, empowerment. (Sheard and Kakabadse 2004:13).

Lofland and Lofland (1995) included in their definition, that a *group* was composed of members who consider that they were a group. They noted that a group is *"A few (up to a dozen or so) people who interact with some regularity over an extended period of time and who conceive of themselves as a social entity (a "we") form a social group. Informal leisure and work groups, cliques, networks, and families are prime examples"* (Lofland and Lofland 1995:103).

Writing about group-dynamics from a strictly psychological point of view, Luft (1984) believed that a group must meet four criteria. These were "*1. Some interaction must take place. 2. Some purpose or goal must be shared. 3. Some differentiation of behavior or function must begin to emerge. 4. There must be more worth or value in being within the group than in being outside of it*" (Luft 1984:7). He also described the group phenomenon of habeas emotum. This is a "*quasi-legal metaphor about fair and just transactions in interpersonal relations. Every person, every group member has a right to his or her feelings and a responsibility to others for the same right".* He believed that in practice, it *"is closely tied to a group's ethics"* (Luft 1984:154).

The Mayer (1992) committee report is one of the relatively rare uses in the educational literature of the word *team*, rather than *group*. They report that one of seven key competencies of students of post compulsory vocational education and training was *working with others and in teams*. They asserted that this competency is *"essential to all work and adult life"* (Mayer 1992:5). It would also be equally important however, for a fulfilling and meaningful childhood and the same for adult leisure, points not addressed in the Mayer text.

From these it is clear that there is no single unifying view of the definition of words group and team. For further examples that show that this difference in the definition may be more than just mere semantics see Strauss (2001).

3.2 Groupwork, group work, group and peer learning, and group assessment

As previously mentioned in section 2.7.4, groupwork experience can be greatly beneficial to undergraduates when they begin their careers (see for example section 3.5 and section 3.5.1 in particular).

3.2.1 Groupwork and group work

In this thesis the terms groupwork and group work are synonymous.

A group project, (or in this thesis, groupwork and group work,) has been broadly defined by some authors as

> "an assignment that requires two or more individuals, interacting and independent, to come together to achieve specific objectives"

(Young and Henquinet 2000:56; Kench et al. 2009).

G*roupwork*, (or group work,) as applied to higher education, has also been defined by Morris and Hayes (1997) as,

> *"where students are required to complete a small-group project … as part of the assessment for a unit".*

(Morris and Hayes 1997)

Strauss (2001) cited Johnson, Johnson and Smith 1998a: 25 as describing group work as:

> *"a technique that 'involves students working together in small groups to accomplish shared learning goals and to maximise their own and each other's learning'"*

(Strauss 2001:55)

These definitions are clear and may or may not lead to summative assessment. The implied or explicit purpose is to improve learning and to enhance the skills involve in working as part of group work.

3.2.2 Group and peer learning

While group learning will always involve a group, peer learning, as noted by Boud (2001) could be in either a group or 1:1. In his introduction to Boud et al. (2001:6), he noted that *"Peer learning need not be primarily about working in groups"*. In the 1999 paper, he noted that

> *"In the context of this paper, peer learning refers to the use of teaching and learning strategies in which students learn with and from each other without the immediate intervention of a teacher."*

(Boud et al. 1999:413-414)

Peer learning has an important role in education for both the teacher and the learner. The topic can be explained peer to peer by language that is more contemporary and by using examples of shared experiences, rather than by a professional teacher perhaps applying more abstract academic examples that may be outdated or less relevant to the learner(s). It has also been said that in order to test one's own grasp of a topic, we should try explaining it to someone else, indicating the importance of peer learning for the peer teacher.

3.2.3 Group and peer assessment

It has been observed that peer assessment *"involves evaluating the product of the group work and assigning marks to students" (Strauss 2001:55)*. In the same paper, Strauss cited several groups of authors writing about the problems of group assessment, including the Lejk et al. observation on the problematic nature of group work assessment mentioned elsewhere (e.g. section 3.9.1). The citations also included Morris and Hayes (1996), mentioned elsewhere (in section 3.10), and one from Watson and Marshall (1995) that *"no prominent researchers currently advocate the use of group assessment as the sole means of evaluation ".*

Other researchers have noted that *"Based on these preliminary findings we conclude that peer assessment can be usefully and meaningfully employed to factor individual contributions into the grades awarded to students engaged in collaborative group work"* (Cheng and Warren

2000:Abstract). This does however contradict an earlier finding of theirs reported by Burd et al. who questioned the "*use of students in the assessment process"*, that *"The results of Cheng & Warren [2]* [1999] *have concluded that student peer assessments are not sufficiently reliable to be used to supplement teacher assessment"* (Burd et al. 2003:56-57).

3.3 Practise effect

The practise effect applies to most aspects of human endeavour. It is not unique to GSA; however, it could be important to it. (Also, see section 8.1.2). The effect has been described as *"gains in scores on cognitive tests that occur when a person is retested on the same instrument"* (Kaufman 1994) for example. Although Kaufmann restricts his description by including the caveat retest *on the same instrument*, he pointed out that the gains are purely due to the candidate's previous experience of the assessment. They happen without there being any feedback from previous tests. Improvement in scores due to the practise effect also has nothing to do with any improvement due to maturity. Kaufman notes that it is a *"systematic, built-in error"*, and that it is *"associated with the specific skills the test measures"*. In addition, they *"relate to the test's psychometric properties"* (Kaufman 1994). He insisted, *"These effects relate to the test's psychometric properties, and must therefore be understood well by the test user as a specific aspect of the test's reliability."*

Kaufman also pointed out that when the time between test and retest is short, candidates would be likely to remember specific test items. These could include *'picture puzzles, arithmetic problems, or block designs'* (Kaufman 1994). They may also be able to recall how they attempted the questions. This would exaggerate the assessment results, he noted.

When Field discussed systematic variation in repeated measures study design, he noted that the practise effect was one of *"The two most important sources of systematic variation in this type of study"* (Field 2009:17). (The other was the boredom effect.)

3.4 The ubiquitous nature of group work and GSA in higher education

There is a large body of literature attesting to the ubiquitous nature of group work in higher education. Burd et al. for example, have postulated that *"Group work projects are one of the few unifying features of universities"* (Burd et al. 2003: Section 1 Introduction). It has also been noted that

> *"Following a lengthy period of reluctance to exploit its possibilities, group-work is now firmly established in higher education".*

(Thorley and Gregory 1994:19)

Additionally, *"group projects are increasingly common in colleges and universities"* (Maranto and Gresham 1998:1), and when Michaelson contributed to the debate, she noted that:

> *"Group learning has been promoted in higher education, and education in general, for a variety of reasons over many years. …*
> *Group learning is advocated by those who see it as preparing students for the world of work by introducing teamwork".*

(Michaelson 2004)

Another researcher was more specific; for him *"Student group projects are a ubiquitous feature of business education"* (Bacon 2005:248). Others have suggested that *"The proliferation of projects using student teams has motivated researchers to examine factors that affect both team process and outcomes"* (Barr et al. 2005:81). This widespread adoption of group working must also include the adoption of the summative assessment of it (also see section 4.5.6).

According to Morris and Hayes (1997), groupwork and group assignments were common practice in many HEI's. They found that it was *'heavily adopted'* in the areas of Marketing, Management and Human Resource Management at Edith Cowan University (Western Australia) and to a lesser extent in some other unidentified disciplines. Several years later, Brown et al. (2006) acknowledged the advantages that groupwork (and therefore GSA) may bring to student learning and employability. At the same time, they included a warning regarding *the challenges* of *collaborative work.* They stated that

> *"a pedagogy that optimises students' academic development is likely to be beneficial to the development of their employability. Such pedagogy will encourage student engagement, amongst other things, by … involving collaborative work where appropriate (**notwithstanding the challenges this introduces regarding high-stakes or summative assessment**)".*

(Brown et al. 2006, emphasis added)

More recently, Grajczonek (2009) has indicated that:

> *"As academics' workloads continue to increase, particularly for those who teach large cohorts such as in undergraduate courses, **the practice of group assessment** [i.e. GSA] **becomes increasingly attractive**".*

(Grajczonek 2009:156, emphasis added)

3.5 Rationale for GSA practice

The rationale for GSA drawn from the literature listed in Table 3.

Table 3: Reasons for GSA use

1	It is what employers want
2	It teaches generic group working skills
3	It is an effective teaching and learning tool
4	It allows larger, more meaningful and realistic projects to be researched
5	It allows better utilisation of scarce resources, including marking

The reasons are addressed separately below.

3.5.1 It is what employers want

Many authors, who have published an opinion, agree that an ability to work in groups is one of the attributes that employers want from new graduate employees although there does appear to be some dissention. (Cohen and Mullender 1993; RSA 1995; Harvey et al. 1997; Leckey and McGuigan 1997; Brown et al. 2006; HEA 2006)

The Higher Education Academy (HEA 2006) reported a broad consensus among employers on the attributes that they expect to find in graduate recruits. They included being able to work in a team. Two years previously, Nicholl and Alexander (2004a) had made the same point. They had noted that the qualities employers demand when recruiting and selecting new graduate staff included *"effective communication and teamworking"* (Nicholl and Alexander 2004a:156). At the same time, they were insistent that, *"computing students tend to find group work difficult"*.

On the other hand, a Royal Society for the Encouragement of Arts, Manufacturers and Commerce (RSA) report from 1995, chaired by Sir Anthony Cleaver, had been less certain that employers really knew what they wanted in their graduate recruits (Cleaver 1995). The report has also been cited by Harvey et al. (1997). It suggested that employers were uncertain about either their current or future attributes needs in their graduate recruits:

> *"a recent study, Tomorrow's Company, suggested that employers are not sure what they really need, or will need in the future and have paid little heed to wake-up calls in the past".*

(Cleaver 1995; Harvey et al. 1997: Chapter 1)

The Cleaver (1995) team analysed interview data and concluded that employers wanted a range of *personal and intellectual attributes.* These were in addition to subject knowledge. They further concluded that there was a difference in skills between what employers want and what HE supplied. In particular, in their view, so-called generic skills and social skills were lacking in graduate skills sets. (Also, see next section 3.5.2). They also observed that the list of employers' desirable graduate attributes seemed to be growing. They asserted that this growth presented a challenge to both students and academics. It seems unlikely that the list of desirable graduate attributes will be any shorter now.

Wolf has also summed up what industry thought it wanted from its graduate recruits prior to 2002, as well as expressing her doubt that core skills could be assessed adequately. Her paper

concerned assessing core skills in general, rather than those of group-working in particular. She explored this from the point of view mainly of YTS (Youth Training Scheme) and NVQ, post-16 students rather than HE students. She noted that *"The "Core Skills" project was dogged by two fundamental problems which, we would argue, are inherent in any attempt to assess and accredit core skills in a free standing way".* These were, she noted, first *"the necessary impossibility of decontextualising statements about core skills with any meaning"*, and second, *"the contingent, but nonetheless overwhelming impossibility of assigning levels in a consistent and comparable fashion",* (Wolf 1991:194). In a later paper, she expressed two additional assessment concerns. First was for the dilution effect that the expansion of student numbers might have on resources. Second was qualification inflation (Wolf 2002).

More recently it was reported that: *"In our recent survey of 13,000 employers, over half felt the education system was failing to equip people with the skills necessary for today's workplace"* (Kingston 2005: xxiii).

On the other hand, Bennett et al. (2000) noted that there might seem to be a close consensus among employers regarding the skills they require in their graduate recruits. They observed that the actual skills required would be context specific to the job tasks and *"Thus, whatever skills training new graduate employees have had in their higher education, it is likely that they will lack the propensities for the varied, yet contextualized, communication demands of their company"* (Bennett et al. 2000:172).

The evidence that it is what employers want seems rather unclear. Despite the Kingston (2005) and HEA (2006) surveys, there seems to be little definitive evidence of what attributes employers think they want graduates to bring to their organizations. Some employers may simply want students knowledgeable in their degree discipline. They may believe that their organization's graduate training or induction programme will suitably instil the required so-called generic skills in their new graduate recruits. On the other hand, it seems likely that others may want their graduate recruits to contribute to their organization immediately. They may want them to fit into a project group without additional training, in which case, their additional groupwork experience may have been an important reason for their selection for employment.

3.5.2 It teaches generic group working skills

Generic skills includes for example working with others, communication, leadership, problem-

solving and IT skills (Hyland 1997; Wolf 2002; Knight 2004; Ford 2007). Also, see the Candy (1995) qualitative learning model in Figure 3 below. They are referred to by some authors as a generic skill set and by others as generic competences, or transferable, soft, key or core skills. The Dearing Report (1997) used the phrase *key skills.* They have also been referred to as professional skills (Harpe et al. 2000). An important part of the published rationale for using groupwork on student degree programmes has been that it will enable students to learn these skills or competences. In addition to the authors mentioned above, several others have also raised this point, (Scriven 1983; Candy et al. 1994; Garland 1996; Holmes 1998; Drury et al. 2003).

Some authors have asserted that students would achieve their generic skills competence from a practical standpoint rather than from a theoretical one. They would derive them through exposure to and engagement with the practicalities of group-dynamics (Leathwood et al. 1999: introduction). Moreover, other authors assert that, *"groupwork skills are among the most important generic attributes students should develop during their studies as preparation for the world of work"* (Drury et al. 2003:1).

Another author who asserted that teamworking is one of the personal transferable skills was Candy (1995). He placed it under the umbrella of *learning ideas and skills of quality*. This means learning outcomes which are valued by and worthy of higher education. His qualitative learning model, which includes *'teamwork'* as a *'personal transferable skill'*, is shown in Figure 3.

Figure 3: Qualitative learning model (Candy 1995)

Knight (1995) was another dissenter. He suggested that these interpersonal skills might not be quite so generic, or transferable. He also extended further warnings, in respect of assessment as a student motivator, which are raised in the next chapter (section 4.5.6).

The introduction to Archambault (1965), warned about the difficulty of adequately defining key skills echoing Wolfe's point mentioned earlier:

> *"There is also considerable agreement among several of the authors on an analogous notion: that teaching for generic skills is, at best, highly suspect, since these are difficult, if not impossible to define".*

(Archambault 1965:12)

If this is true, in modules where their acquisition is one of the teaching aims, there will be similar difficulty in teaching to achieve this. They will also be just as difficult to assess.

Despite the overall lack of clarity noted above, the HEA (2006) reported a broad consensus among employers on the attributes that they expect to find in graduate recruits which include the ability to work in groups. Of course, student groupwork cannot be identical to workplace teamwork and they may have just taken student groupwork experience as a best compromise.

The Drury et al. (2003) use of the more tentative and conciliatory word *should*, in their comment that *'groupwork skills are among the most important generic attributes students should develop during their studies as preparation for the world of work'*, rather than something more positive and reassuring gives their observation a note of caution.

In the literature, it is unclear whether the dissenters against the existence of generic or transferable skills had doubts over the concept or merely disagreed with the terminology. For example Holmes' (1998) use of the phrase *personal competences and capabilities* rather than *generic skills*.

3.5.3 It is an effective learning and teaching tool

Part of the rationale for GSA practice is that it is an effective learning and teaching method. There is some disagreement in the literature over whether or not it is effective for students. On the one hand, Barfield (2003) asserted that its effectiveness as a learning and teaching tool was *"widely accepted"* by HE researchers (Barfield 2003:355). On the other hand, at least one researcher has doubted its efficacy and even asserted that it *actually inhibited* topic content learning (Bacon 2005:248). The issue would seem to be with the definition of the term *effective*. Michaelson (2004) was even more sceptical about the effectiveness of GSA. She expressed no more than a mere *hope* that groupwork would result in deep and active learning, compared to learning that is shallow and passive. (Michaelson 2004)

Assessment is most effective, when it reflects an understanding of learning as multidimensional

and integrated (Jaques 2000). Groupwork is one of the approaches advocated for this purpose.

There were several papers found of specific aspects of group assessment practice, (e.g. Isaacs 2002; Burd et al. 2003; Nicholl and Alexander 2004a). None however has reported randomised control trials (RCTs), or systematic review (meta-analyses) studies of the impact of GSA as a learning and teaching tool.

3.5.4 It allows more meaningful and realistic projects

Groupwork and GSA practice may provide an opportunity for students to engage in projects of real world realism. In this context, this seems to mean appropriate pressure on the project group to succeed, especially in terms of level of work, delivery deadlines and reward for effort. The greater the numbers of students in each project group, i.e. the resource, the more far reaching might the research be for their module project (Isaacs 2002; Nicholl and Alexander 2004a). Isaacs, for example also affirmed that GSA could benefit both students and staff. He added the caveat *"only if it is carefully planned"* (Isaacs 2002:7).

One of the problems with this fourth part of the rationale for practice is in the detail, i.e. in the definition of the terms *meaningful* and *realistic*. For example, few GSA module projects were found where the project was for an actual commercial client, and there was only one in the study data. It was a physics industrial project module. The GSA mark for the group component contributed 85% of the individual's marks for that module. Each student received the same mark for the group task. The task was to produce an artefact, an instrument of some kind. They negotiated the specification with their real world client to help to resolve their client's real world problem (Myers 2007).

Another example, this time of a large group size, was an applied social sciences module. The whole seminar group of fifteen, or sometimes more, students all contributed to a GSA poster (Boyne 2007). The mark from this assessment item contributed 20% of the individual student's module mark.

By allowing the opportunity for more meaningful projects, GSA modules may also be part of the accreditation requirements of some professional institutions. Example one, a substantial group project is one of the requirements of degrees accepted for entry to the British Computer Society, (BCS 2007), (also, see section 2.7.3). Example two, the Royal Pharmaceutical Society of Great Britain require schools of pharmacy to *'seek to develop students' skills of team working'* (Gidman et

al. 2008).

3.5.5 Resources may be used more efficiently

GSA may have become increasingly attractive to teaching staff as a response to increased academic workload and the so-called *'massification'* (e.g. Baty 2009) of higher education. It could also be a more efficient use of diminishing resources, i.e. funding per student, and help to deliver equivalent programmes to increasing class sizes (Grajczonek 2009).

Garland, in her subsection on reasons for setting group work, benefits for staff, found that:

> *"The majority of respondents agreed group work saves time when marking students work, and groupwork is less expensive to set up, supervise and equip than projects undertaken individually".*
> (Garland 1996:225)

She added, *"Taken together … replies suggested respondents believe group work to be less costly both in terms of human and physical resources"* (Garland 1996:225). She was emphatic about the time saving aspect of GSA. Gibbs explained her first point more cautiously when he noted that: *"Group project work appears to offer … the possibility of greatly reduced marking loads, especially where lengthy and complex products emerge from project work"* (Gibbs 2009).

Perhaps Leathwood et al. covered this particular *efficient use of scarce resources* point best when they noted:

> *"On a pragmatic note, the current emphasis on group work may also be a response to rising student numbers and the increasing pressures on academic staff".*
> (Leathwood et al. 1999)

On the other hand, Nicholl and Alexander (2004a) warned that increased time for group support and supervision could replace that saved by reduced marking.

This fifth reason claimed for GSA practice, that resources might be used more efficiently, is too general an assertion. For example, did the majority of Garland's (1996) respondents agree that this applies at all times, to all groupwork? Were there any dissenters, if so, what did they have to say? Were there any reductions in efficiency or coordination losses (Bacon et al. (1999)) due to the extra resources that students might need to prepare for groupwork, e.g. allowing time to practise group-dynamics skills? Gibbs (2009) was more pragmatic in his on-line article.

In some circumstances, GSA may provide some educational institutions with opportunities for greater efficiency in the use of some teaching resources. This includes efficiencies with marking

resources required for some degree programme modules. Personal experience, participant observation and oral evidence from peers, from both sides of the andragogical fence suggest that this may be an important contributor to the rationale behind higher educational institutions adopting GSA.

Module teaching resources could be either human or physical. They both depend on adequate funding. The minimum usually required for module delivery are staff and a suitable teaching venue. Distance learning modules will also need a communication system. In addition, teaching staff involved in summative assessment also need adequate time to assess and score students scripts and GSA products. Some modules that use GSA may require more resources. Specialist programmes may need specialist equipment, teaching staff training or assessors. In some disciplines, specialist laboratory equipment could require specialist design, manufacture, maintenance, and/or handling and storage. They could also require special security precautions, e.g. X-ray laboratory equipment. Modules like these may be too costly or complex to be viable without the efficiencies associated with undergraduate group work. Summative assessment would then be an important part of the group work.

Efficiency, all else being equal, is a perfectly sensible reason for pursuing or amending any course of action, not just assessment practice. Not to do so would mean that the practice was inefficient. There does not always have to be an andragogical reason to change to GSA. Efficiency is reason enough. Thorley and Gregory (1994) were wrong to issue their rider in relation to making sure that it is not used only to save staff time because this is an excellent reason for change. Thorley and Gregory (1994) made rather an odd comment. If there is no student downside to the groupwork, why did they not want it used purely with the intention of saving staff time? This would seem to be a most admirable aim. There was little discussion on this. It would release finite resources for other purposes without detriment to the existing module. What they may have meant was; not used *only* with the intention of saving staff time, without any andragogical concern, but this was not clear.

It might literally be more efficient to mark one group individual summative assessment item than one from each student. This simplistic view does not take into account the lifetime resources that may be required. In addition, for fairness, it should only be applied if e*very student* has the same level of experience of GSA and group-dynamics or an equal opportunity to prepare for it.

Resources will be needed for GSA teaching and marking, in a similar manner to IISA modules. In addition, other aspects, unique to GSA, will need additional resources. These include, time for students to practise the interpersonal skills that GSA requires. They will also be needed for formative feedback on student attainment in these key skills. In particular, there will be a need for resources for the additional queries that may arise because of the novel assessment method. This may even extend to general and individual reassurance. Perhaps even some academic adjudication will be needed on group-dynamics issues like intra-, or in more extreme cases inter-group, disputes.

Any group project has the potential to be much larger and more complex than a similar individual assessment in an equivalent IISA module. It may take proportionately longer to assess group work than several smaller, simpler IISAs from the same number of students. Consider, for example, a group essay. An individual undergraduate coursework assessment essay might have a maximum word count of say three thousand words. It would be impractical to set the word count of a group essay in proportion to the number of group members and setting a restricted word count would be problematical. The need for the group to produce a compact essay with a disproportionately low word count could increase the complexity. In addition, early in their study programme, students may not always be concise in their writing. Overall, this could require more time to mark, rather than less.

3.6 How student GSA groups are formed

The three basic methods used for group membership selection will be described and reviewed in section 3.6.1. Issues of group size will be in section 3.6.2. The groups are often quite small, most commonly being in the range of four to six members. The final section, 3.6.3, will review some of the literature on students' so-called learning styles.

3.6.1 Group membership assignment methods

There are three basic ways of assigning students to groups (Bacon et al. 1999:468). The choices are self-selection, random (or more likely pseudo-random) assignment, or teacher assignment. The method selected should depend upon the learning aims of the module. The literature is reviewed below in sections 3.6.1.1, 3.6.1.2 and 3.6.1.3.

3.6.1.1 Self-selection

Bacon et al. (1999) have suggested that there is a sound andragogical reason for using self-

selection. They noted that some researchers recommend it because it may offer higher initial cohesion. This may link to higher group performance. They argued that this was important because student projects often have a relatively limited lifespan. Initial cohesion of self-selected groups may help them to become productive more quickly. Self-selection may also encourage students to take ownership of the group project. Students might then be able to manage interpersonal tensions more successfully. It has also been noted that students would often request to work with those with whom they have worked previously, and additionally that:

> *"Self-selection is not without problems however, including the tendency for self-selected teams to be overly homogeneous, and thus not offer the advantages that some diversity may provide".*

(Bacon et al. 1999:468)

Group membership self-selection can result in the lack of a more appropriately varied group skills set. (Tuckman 1965; Mello 1993; Bacon et al. 1999)

3.6.1.2 Random assignment

Random allocation, or perhaps more often pseudo random allocation, of students to GSA module project groups may seem fair; but as Bacon et al. have suggested, assigning groups like this is

> *"just as unfair as randomly assigning grades - each student would have the same probability of getting an A or an F, regardless of their abilities or efforts".*

(Bacon et al. 1998:69; Bacon et al. 1999)

Theirs may seem to be an extreme viewpoint but as they explained:

> *"Each student begins the class with the same chance of working with every other student, but due to the random nature of this approach the final team assignments can be quite unbalanced in terms of skills, diversity and general ability. Random assignment is also not likely to generate teams with a useful combination of skills, or create groups of students who want to work together. We suspect that whereas some randomly assigned teams would, by chance, end up with a desirable combination of students, others would certainly not, and therefore random assignment would not generally be associated with good team experiences and may be associated with bad experiences".*

(Bacon et al. 1998:69)

Hackman (2004) reminded us that psychologists have long recognized that the best predictor of future behaviour is past behaviour. The article was on the subject of the selection of leaders in commercial companies and the ways in which the attitudes and behaviour of *team* members complicate the chances of a team's success. It was not about student groups. His belief in the existence of what he referred to as team-destroyers is also noteworthy. These were

> *"people who will undermine any team you put them in. Such people may be so unskilled in working collaboratively with other people, or so individualistic in their focus, that they should be invited to make what may be an excellent contribution to their organization as solo performers".*

(Hackman 2004)

This supports the Thorley and Gregory (1994:178) observation cited at the start of this thesis, that

some students may work better alone. Hackman went on to reassure us that there are far fewer such people than we might think.

3.6.1.3 Teacher assignment

Student assignment to groups could be made explicit by the teacher of a module. It could also be from a pre-determined algorithm or it could even be context dependent. Alternatively, it may be that its aims and learning outcomes allow and encourage negotiation of the method of group allocation, between teacher and students. Also, see example in section 2.7.1.

Bacon et al. (1999) suggested that teacher assignment is the least popular method of group allocation, and can be difficult to implement. Potential hazards could be, for example, the possibility of future claims that the group allocation method was unfair, from students who later did not receive the (higher) marks that they thought they deserved.

Tutor selected grouping methods can also be adjusted to address local situations. The methods could balance or select the group based on gender for example. Self-selected teams may have an inadequate skill set, unless measures are in place to moderate it.

3.6.2 Group size

Laughlin et al. (2006) studied 760 students at the University of Illinois and concluded

> *"groups of size three are necessary and sufficient to perform better than the best of an equivalent number of individuals on* ***intellective*** *problems".*

(Laughlin et al. 2006:650, emphasis added)

The Laughlin et al. (2006) study subjects were university students in the United States. They were studying at a similar level to those in this present study. It may be useful to note however that Laughlin's students also received course credit for participating in the study, which may have influenced either their decision to participate, and/or the study outcome.

Olson (1965) was another author who also supported the hypothesis that small is best. He noted that *"Small groups will further their common interests better than large groups"* (Olson 1965:52). In addition, he had earlier shed further light on group size issues with his insightful view that

> *"The difficulty of analysing the relationship between group size and the behaviour of the individual in the group is due partly to the fact that each individual in the group may place a different value upon the collective good* [product or service] *wanted by his group".*

(Olson 1965:22)

Bacon et al. (1999) found agreement among authors about group size. They advised that they should be kept as small as possible, to reduce the opportunity for social loafing, also referred to as free-riding, see section 3.9.7, in student groupwork literature. They also noted that disharmony within the group tends to increase, while attainment remains static, with increased group numbers. As group size increases, communication between group members will also become more complex. They called this coordination loss. Bacon et al. also asserted that the student perception of the groupwork experience improved with increasing group size. In a larger group, there is less likelihood of a student being discovered social loafing. The miscreant's contribution or lack thereof will be more anonymous. In addition, there will be less risk of serious repercussions if they are unmasked.

In a later work, Bacon (2005) observed that in the context of peer learning there is no optimum group size. On the other hand, he noted firm evidence that a group of two was enough to reap the benefits of peer learning if the task required both interdependence and individual responsibility. Another author noted that *"As the size of group increases, so its characteristics change"* (Jaques 2000:7). He illustrated this in his figure 1.1, reproduced in Figure 4.

More cohesion

Number of members	*Changing characteristics*
2–6	Little structure or organization required; leadership fluid.
7–12	Structure and differentiation of roles begins. Face-to-face interaction less frequent.
12–25	Structure and role differentiation vital. Sub-groups emerge. Face-to-face interaction difficult.
25–?	Positive leadership vital to success, sub-groups form; greater anonymity. Stereotyping, projections and flight/fight occur.

More tension

Figure 4: Group membership number effect (Jaques 2000)

3.6.3 Learning styles

Learning style is a concept challenged by some educationalists, but commonly used (Coffield et al. 2004; Kingston 2004; Coe 2010) and it is another method of forming student groups according to their self-reported styles.

There are many learning styles models. Evans (2003) found evidence of over 125 different models of cognitive and learning styles. She found little consistency between the methods used to differentiate them. She warned that:

> *"This leads to a consideration of whether this reveals the complexity of cognition or whether the various styles are simply different conceptions of the same dimension".*

(Evans 2003:14)

In addition, Coffield et al. (2004) reviewed, in their words, 13 major examples. At the start of their substantial report, they were equally critical of the learning styles concept.

> *"The commercial gains for creators of successful learning styles instruments are so large that critical engagement with the theoretical and empirical bases of their claims tends to be unwelcome".*

(Coffield et al. 2004:2)

They asserted that the learning style theorists claim that they *measure* the learning preferences of students is false. They derive it from the students self-reporting. This was usually by questionnaire. Like Evans, they warned that critics of learning styles also *"dispute the objectivity of the test scores derived from the instruments"* (Coffield et al. 2004:46). They asserted that they were mainly those who advocate qualitative rather than quantitative research methods. In their summary tables of the methods they reviewed, they also noted that some of them either looked *'promising'* or had at least some *'potential'*. The Allinson and Hayes Cognitive Styles Index has *'the best evidence for reliability and validity'.* The Entwistle Approaches and Study Skills Inventory for Students were *'potentially useful.'* Herrmann's Brain Dominance Instrument offered *'considerable promise'*, as did Jackson's Learning Styles Profiler. On the other hand, the Coffield et al. view of the Honey and Mumford Learning Styles Questionnaire was that it needed to be redesigned. Of the Myers-Briggs Type Indicator, they observed, *"it is still not clear which elements of the 16 personality types in the MBTI are most relevant for education"* (Coffield et al. 2004:31). (Belbin 2004; Moreland 2004)

Some authors have referred to methods of assigning students to project groups as *"Attempting to 'engineer groups' according to personal characteristics"* (Huxham and Land 2000:17-18). In addition, Huxham and Land found *'no significant difference'* in assessment performance between groups randomly selected, and those based on the Honey and Mumford method. They also noted that, in their published advice most authors advised attempting to *'engineer groups'*. Despite this, they found that the other two methods, self-selection and random, or pseudo random selection were more common. Their questionnaire identified four learning styles: 1) Activist; 2) Reflector; 3) Theorist and 4) Pragmatist. They were based broadly on the Kolb learning cycle. This compares, for example to the sixteen learning style categories used by the Myers-Briggs Personality Type Indicator (Amato and Amato 2005). (Also, see section 8.2.11, for additional discussion on the

MBTI psychometric evaluation method.) Some researchers had reservations on their questionnaire (Zwanenberg et al. 2000; Duff and Duffy 2002). Coffield et al. wrote a detailed discussion of '*The objections to learning styles'* (Coffield et al. 2004). Equally, Duff and Duffy were also sceptical of methods of group forming.

3.6.3.1 Myers-Briggs Type Indicator

Some authors have asserted that it would be useful for students to have an understanding of different personality types. They would then have a vocabulary and a knowledge base *"for discussing conflicts and for balancing preferences in team assignments"* (Amato and Amato 2005:42). They were discussing the Myers-Briggs Type Indicator (MBTI). The four MBTI personality types continuum is in Table 4.

Table 4: Myers-Briggs Type Indicator (MBTI) personality types

Continuum	Abbreviated to	MBTI Category
Extrovert / Introvert	E/I	Expecting and receiving energy
Sensing / Intuiting	S/N	Getting information
Thinking / Feeling	T/F	Decision making
Judging / Perceiving	J/P	Attitude to the external world
Amato and Amato (2005) and Schullery and Schullery (2006)		

They lead to sixteen type categories that they gave descriptive names. The one that Myers-Briggs calls the Inspector, for example, is introvert, sensing, thinking and judging. They use the acronym ISTJ. It is apparently the most common category among post-graduate students (Crabtree 2009). Another example is ENTJ, which they call the Executive.

The MBTI questionnaire was completed by this researcher in preparation for the Crabtree (2009) post graduate seminar. How some of the questions can be useful in assessing their MBTI type is difficult to understand. Figure 5 is the screen dialogue box content for the MBTI question number fifty-nine.

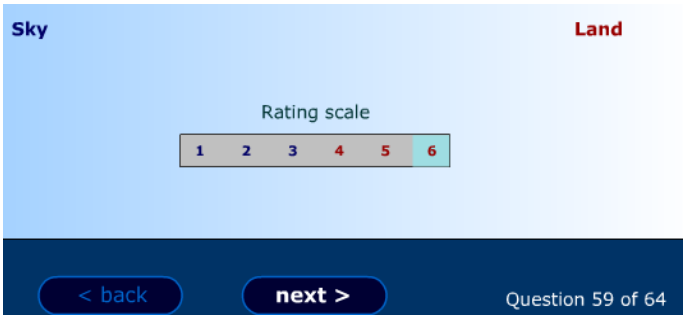


Figure 5: Part of the MBTI question 59 screen dialogue

The question was asked just as it appears in Figure 5, i.e. '**Sky Land**'. The instruction for this

range of questions was to indicate the word that most appealed to the candidate and which fitted how they see themselves. This latter addition also made it a multiple instruction because they were asked to respond to two separate concepts. Although there may be solid evidence that, for example, people who choose *"land"* have a particular 'learning style' or are of a particular type, this was not discussed. The inclusion of this item seems to cast doubt on the validity of the MBTI test items in the manner of the Hardy/Twain *thirteenth stoke of a clock.* It is *"not only false by itself, but casts grave doubts on the credibility of the preceding twelve"* Light and Pillemer (1984). In addition as mentioned earlier, the data that the MBTI is based on is self-reported, putting its reliability further in doubt.

3.6.3.2 Belbin management types

The Belbin experiments, *The Management Game*, were a series of interventions with management student groups (Belbin 2004). He concluded that group members have preferred roles depending on their personalities and abilities. He found that in order to succeed, management groups needed the right mix of key team role types, which were devised and defined by him. The names of these team roles was often more idiosyncratic than descriptive. Plant and Shaper were two examples. Other names however were more self-explanatory, e.g. Chairman, Company Worker, Team Worker, Monitor-Evaluator, Resource Investigator and Completer. He noted that a group composed entirely of clever people of the same key team role would be likely to fail. This seems hardly a finding of note. He called this failing of a group of clever people the Apollo syndrome (Belbin 2004: Chapter 2). The Apollo team were a team of clever people but it did not have an optimum mix of abilities and attributes. Their performance often compared unfavourably with other groups of less able people. Belbin suggested that one possible reason for the Apollo group's failure was *"the tendency of clever people to overrate cleverness"* (Belbin 2004:17). He suggested that the Apollo team would need an *Apollo Leader* as Chairman. They would need slightly different attributes to those of the standard role of Chairman. He commented, *"In human affairs nothing should be taken for granted. … The Apollo team finished last"* (Belbin 2004:10). Similarities between the Belbin Apollo team roles and outcome, and the behaviour and outcome of some student summative assessment groups may explain the failure of some student groups. Belbin explained that, "*The danger in many groups is that individuals strive competitively to make their voice heard on any matter that comes up; no one is interested in heeding anyone else"* (Belbin 2004:65).

3.6.3.3 Heterogeneous versus homogeneous personality groups

Schullery and Schullery (2006:549) reported on a classroom study in a business college. They

studied the benefits to students of working in one of two types of personality groups. The two types were homogeneous and heterogeneous. One group was made up of compatible personalities and the other was made up of complementary personalities. Students were divided into 102 groups of from two to six members. They used the MBTI instrument to establish personality types within the groups, (see section 3.6.3.1). One of the four outcomes they used to assess the study was student group projects grades. The students were assessed individually, they *"received individual grades based on their portion of the report"* (Schullery and Schullery 2006:546).

They used four different variables to assess outcomes at both the individual student, and the group levels. Three of their variables were student orientated: 1, self-perceived improvements on a variety of group skills, 2, students overall satisfaction with their group experiences, 3, student grades on the group projects, and the fourth was the *'instructor's'* perception of how well the group functioned.

Group heterogeneity, they suggested, was possible by selecting the individual members' attributes and characteristics for that particular outcome. They noted that

> *"a teacher can situate her students to maximize their grades and satisfaction with their groups by assigning groups that are heterogeneous with respect to argumentativeness".*

(Schullery and Schullery 2006:554-555)

They were unable to draw firm conclusions from their study.

> *"There is no simple answer to the research question; participation in heterogeneous groups is associated with advantages and disadvantages for students".*

(Schullery and Schullery 2006:554)

3.7 The impact of GSA on student marks

Rothwell (2002) cited Lejk et al. (1999) as an example of the impact that GSA might have on students overall marks. This was a specific example of a module where groups were tutor allocated according to their ability.

> *"Students who had done the best on the first two* [IISA] *tests averaged 11 percent lower marks if they were assigned to mixed groups than those who had previously done well and were assigned with others who had also done well. Students who had done poorly on the first two tests scored an average of 12 percent higher when they were assigned to mixed groups than those who were assigned with others who had not initially done well. The implications of these findings are clear, "the method by which a group is formed seems to have an effect upon the performance of the group"".*

(Lejk et al. 1999:13; Rothwell 2002)

Downie (2001) has hypothesised that, due to its emergent properties, or gestalt in psychology, the

whole of the output of the system or group, is greater than the sum of the output from its individual members. She noted that in group work projects, this

> *"may also result in a lower range of marks with higher means because the product of group work can often be better than individual submissions".*

(Downie 2001:7; Johnson 2005)

The finding is supported by other researchers. It has been noted that *"Cooperative groups perform better than independent individuals on a wide range of problems"* (Laughlin et al. 2006:644).

The phenomenon was also reported by several of the module convenors who were interviewed as a preliminary to this study (also see section 5.1.4).

3.8 An unverified claim

One university web page reported a similar finding, from other researchers, to one from this study on the impact of GSA. This was that,

> *"Some research indicates that low achieving students tend to achieve higher than usual scores when they work in groups but that high achieving students tend to receive similar grades to those they receive when working alone" (Exley and Dennick, 2004).*

(Wollongong 2006)

It indicated that GSA tended not to disadvantage high achievers. It does however confound one of the findings by the Lejk et al. (1999) study cited in the previous section, for example. Their finding was from a Science Faculty cohort. If verifiable, this would have been a very important finding. Also, see Figure 30 in section 7.4.4.2.

A perusal of both of the Exley and Dennick (whom the Wollongong web page cited,) publications found for that year, failed to find this text. Contact was made with Dr. Exley by e-mail in January 2006. Her reply indicated an enthusiastic support for this current project. She was however, understandably unable to recall the text. Additionally, the original University of Wollongong URL is no longer active, (September 2010).

3.9 Disadvantages of GSA

Graham Gibbs has written extensively on using groupwork for student learning and teaching. He has pointed out that because of their flat hierarchical structure and equal membership status, student teams *"are quite unlike management teams"* (Gibbs 1995b:3). He noted that where teamwork training had been introduced in modules, it tended to be based on books and manuals on teamwork in organizations outside higher education. Later he suggested that GSA projects

might disadvantage the more able students. He suggested there was a possibility of them being *"dragged down by poor or lazy students*" (Gibbs 1995b:6). This phrase was also used by Jaques (2000:225). The other disadvantages of groupwork that Gibbs raised included the potential for loss of individual choice. Another was that students might not necessarily learn all the aspects of the task if they only worked on one or two parts of it. He also observed that a student aiming for a first might not appreciate working with others who have ambitions that are more modest. However, he did assert that by the use of various, what he referred to as *assessment devices,* high ability students could avoid these concerns. These *'devices'* were not discussed in any great detail.

On the other hand, his list of advantages included some students welcoming the opportunity to work in teams because they find studying alone socially isolating, although there was no evidence offered to support this. Other advantages, Gibbs observed, include teamwork experience and the opportunity to develop additional learning skills. These experiences and skills were he assured us, *relished* by employers. He also asserted that working in teams allows for more interesting tasks. It also allows peer learning and teaching; and presents a reduced risk of a student *"completely ploughing a project and getting very poor marks"* (Gibbs 1995b:6). He noted that individual work allowed for an assessment of the individual student's performance with a personal piece of work, and that promoted deep learning. He also believed that that approach would mean more marking, and that the assessment task may be boring for students. He compared this to group work that encourages teamwork, maximises available resources, cuts down marking requirement and allows students to share the assessment workload.

The meaning of the Gibbs (1995b) and Jaques (2000), phrase *"dragged down by poor or lazy students"* seems clear although neither of them elaborated on it. One of the conclusions of this study is that the effect they described is due to the GSA method. This outcome is supported by the limited evidence available (Lejk et al. 1999; Hoffman and Rogelberg 2001; Almond 2009).

If lazy students caused the marks of others to be *'dragged down'*, how did Gibbs and Jaques define lazy? Neither of the authors used either the term social loafing or free riding, more commonly found in the literature to describe this situation. Other terms are also used, (see section 3.9.7). Perhaps apparent student laziness could be a symptom of a more serious problem. Gibbs also explained that a student aiming for a first might not appreciate working with others who might have *ambitions* that were more modest. Doing most of the project work themselves, as he suggested

might happen, might then also negate part of the module learning aims. His assertion may perhaps seem a little naïve because it does not appear to take account of the effect of personal interaction, covariance and group dynamics within the group.

The downside to group work, as Knight (2004) saw it, was firstly of the individual becoming subsumed within the group, and secondly, problems with student non-contribution. He summarised the advantages and disadvantages of a group task compared to an individual task. His table 1 (Knight 2004:64) is reproduced here as Table 5.

Table 5: Advantages and disadvantages of individual and group assessment (Knight 2004)

Scenario	Advantages	Disadvantages
Individual work (e.g. essay/report)	• Gives an individual assessment of performance • Is a personal piece of work • Promotes deep learning approaches	• A lot of marking • May be boring for the students
Group work (e.g. presentation/report)	• Encourages teamwork • Maximises available resources • Cuts down on marking • Shares workload	• Individuals get subsumed within the whole • Some students get away with doing little

Michaelson (2004) has suggested that some staff may be resistant to, as she called it, *group-based learning*. She suggested that teaching staff might be anxious about the change in the teacher – student relationship. There could be for example a loss of authority over student marks following the introduction of self- and/or peer-assessment.

The work published by Graham Gibbs and others, on the advantages and disadvantages of groupwork was reviewed above, and the advantages, both real and perceived, were reviewed in sections 3.4 and 3.5 above. The next sections review specific disadvantages of groupwork noted in the literature.

3.9.1 GSA is problematic

In addition to warnings in the HE literature about the ubiquitous nature of GSA (See section 3.4), numerous authors have warned of its problematic nature.

The Morgan (2005a) review of the groupwork literature was inconclusive. It showed that students' experiences were often *"less than ideal"* (Morgan 2005a: Abstract). This is a direct contradiction of the Stephenson assertion (in Thorley and Gregory 1994) see section 3.10, that groupwork *"can also be fun"* although it was almost a decade later. Morgan also found studies that indicated students' experiences could also have a positive effect on their learning. Some authors had

commented on the importance of supporting students. Providing groupwork training would encourage this effect, Morgan noted. In Salomon's terminology, *"Teams do not always function well"* (Salomon 1992:63-4). He added that if the extent of groups' true collaboration was the measure of success then it was rare.

The Stephenson assertion, in Thorley and Gregory (1994), that groupwork *can also be fun* is probably literally true because he included the qualifier *can*, i.e. it might be. He did not elaborate on this; neither did he cite any sources for this assertion.

Leathwood et al. offered this insight. *"Group work can be fraught with difficulties"* (Leathwood et al. 1999: Introduction). They observed that this is the storming stage in Tuckman's *Developmental sequences in small groups,* (Tuckman 1965), also see section 3.9.5. Leathwood et al. also noted Tuckman viewed the progression of group projects as linear, but expressed their doubt over his interpretation of progress. In addition, they asserted that the more frequent reasons for the difficulties of groups were the different levels of commitment and contribution from members. Staff skills and time demands, and groups not having the right mix of skills or training also contributed to group difficulties Leathwood et al. explained.

Some authors have assured readers that *"most teachers and lecturers can attest to the difficulties of using group-work projects"* (Huxham and Land 2000:17). They confirmed that ineffective collaboration was a common issue, and that this could be because *"students lack the very transferable skills which the process of group work is intended to teach"*, furthermore, *"Without explicit instruction in, and practice of, these skills, many groups will not operate well"* (Huxham and Land 2000:17). They also reiterated that providing these extra resources would be time consuming and expensive and so might negate the greater efficiency rationale for group work.

In addition, in the literature, there are similar warnings that *"Groupwork is difficult and complex"* (Heathfield 1999), while Strauss observed, *"Perhaps what emerges most strongly from the literature is that group assessment should be implemented cautiously"* (Strauss 2001:64). It has also been noted, more recently, that any group learning activity is complex and will *"require careful, ongoing facilitation from a skilled instructor",* (Brown and McIlroy 2011:689). Similar warnings can be found elsewhere. (Janis 1972; Sydney 1999; Pitt 2000; Burd et al. 2003)

Sanders (2008), a pseudonym, warned that not all students would warm to group work, nor presumably to GSA. The paper was based on a particular initiative to encourage a class of '*graduate students'* (i.e. post-graduate students) to engage with groupwork. The author reported that they were *resentful* of it. Sanders noted:

> *"I tried to persuade them: "Don't you agree that pooling your resources is likely to guarantee a better end product, especially given that the project requires putting in place so many individual components?""*

(Sanders 2008: emphasis as original)

The author also described telling the students "*cooperation is supposed to confer evolutionary benefits on any species that practices it"*. This was, Saunders noted, in an effort to inspire the students to engage with group work.

Nicolay has also warned that group, or team, working is not everyone's choice. He argued that those teachers who use it *"make it so"* (Nicolay 2002:42). That is, students become part of a team whether they like it or not. In addition, he warned that teachers who use GSA would find themselves spending time reflecting on the process and adjudicating on intra-group tensions and conflicts. In his view they would be

> *"compelled to make sense of the experience, rationalize its appropriateness, and invariably wrestle with the dynamics of the interaction between people who depend on each other to meet personal goals".*

(Nicolay 2002:43)

Perhaps Lejk and his colleagues expressed it most succinctly. *"All in all, assessment of groupwork is problematic"* (Lejk et al. 1999:11).

3.9.2 Adam Smith's 'Invisible Hand'

Adam Smith 1723-1790), in what was probably his most famous work, *An Inquiry into the Nature and Causes of the Wealth of Nations* (1776), noted that each *'producer'*

> *"intends only his own gain, and he is in this, as in many other cases, led by an invisible hand to promote an end which was no part of his intention. Nor is it always the worse for the society that it was no part of* [the individual's intention]*. By pursuing his own interest he frequently promotes that of society more effectually than when he really intends to promote it". (Chapter II)*

(Smith 1976; Joyce 2001:1; Hardin 2003)

As Joyce noted, *"The theory of the invisible hand is certainly persuasive, and its simplicity is also very attractive"* (Joyce 2001:3).

3.9.3 A legal issue

There is also a legal issue associated with GSA. Without careful attention to detail, and transparency and clarity of teaching methods and learning outcome aims, the result of GSA may be

unexpected. It may sometimes even need to be resolved through the law courts. Morris and Hayes demonstrated this extreme with an example that highlights a potential free-rider legal implication. The use of a group project as part of the student coursework resulted in one student originally receiving a module mark of zero for non-contribution. The student successfully appealed the mark. As Morris and Hayes reported in their Australian study,

> *"The appeal was upheld on the grounds that it was specified in the unit outline that the assessment was on a group basis and while* ***admitting to having not contributed at all to the project****, the student argued successfully that he had a legal entitlement to pass as all other members of the group had passed".*

(Morris and Hayes 1997: emphasis added)

There could have been several other known and unknowable external factors contributing to the legal issue; e.g. the phrasing of the module assessment criteria, and how this was communicated, and received, by the student cohort.

3.9.4 Groupthink

Janis, in his 1972 psychological study of US foreign-policy decisions and fiascos, developed groupthink theory from Bion's (1961) group mentality argument. Succinctly put *"Groups, like individuals, have shortcomings. Groups can bring out the worst as well as the best in man"* (Janis 1972:3). Janis posited that groupthink was the suspension of common sense. It was a situation where the individual submits to group pressures, or where individuals

> *"are deeply involved in a cohesive in-group, when the members' strivings for unanimity override their motivation to realistically appraise alternative courses of action".*

(Janis 1972:9)

This means that when the group members become too comfortable in the group, it may underachieve. In such cases, individual members choose not to consider all their options fully, because they do not want to challenge the status quo of the group norms. A degree of internal group conflict, not enough to damage the delicate group-dynamics, may help to produce better outcomes. (Forsyth 2006)

3.9.5 Group lifecycle

Tuckman (1965) called the phases of his original group lifecycle *Forming*, *Storming*, *Norming* and *Performing.* His belief was that a group could not immediately begin the performing phase, i.e. begin work on the task deliverable item. It must first undergo all the other phases in the correct sequence.

Gersick (1989; 1990a) disputed this assertion and suggested a *punctuated equilibrium model* was

more suitable. She had observed eight groups from six different organisations. They included a bank, a hospital and two universities. From her rather short, 2-week, study, she hypothesised that group behaviour followed two separate phases. It had a definite transition point between them near to the mid-point of the time for the project. Another researcher has suggested that; *"another interpretation is that students never really undergo a full and effective 'storming' stage of group development"* (Kates 2002:22). The Kates hypothesis was that they do not wish to disturb the fragility of the group cohesiveness.

When Gersick (1990) disputed Tuckman's (1965) forming, storming, norming and performing work, she was referring specifically to what she called a Task Force group category. A Task Force was one of seven categories of group to which, she asserted, student project groups belonged. In addition, her study was only of two-week duration. If the groups in her study were formed only for that long and if this was the basis for her dissension then perhaps her conclusion was understandable.

In later years, in collaboration with Jensen in 1977, Tuckman added *adjourning* because the lifecycle of a group also include its demise, and it is not over until the final phase is completed. The original *norming* changed to *renorming, adjourning* changed to *mourning,* and it also had *informing* added.

3.9.6 Additional disadvantages of GSA

Students may also be resistant to the group summative assessment because of doubts over the adequacy of the standard of work of their peers. They may also worry that their peers may mark the work of others unfairly by deliberately marking them down.

3.9.7 Deliberate non-contribution to the group effort

In 2006, Forsyth cited Hoffman and Rogelberg as asserting that

> *"Many students avoid group projects where the entire group receives the same grade, because, inevitably, one or more members of the group will not do their share of the work".*

(Forsyth 2006:302) (Karau and Williams 1993; Hoffman and Rogelberg 2001).

One of the methods of dealing with recalcitrant group members, thought not to be contributing a fair share to the group effort, utilizes a soccer analogy and was described in Lejk et al. (1996). The miscreant group member was first to be shown the *Yellow Card*. If there was insufficient improvement, they received the *Red Card,* and they were then excluded from the group.

The deliberate non-contribution to the group effort is known by different authors under different terms, Oakley (2004) referred to students, i.e. US school pupils, who do not do their fair share of the work as hitchhikers and couch potatoes. *She used 'mini-clinics'* as part of classroom activities to resolve these difficulties.

Some authors, on the other hand, have noted that,

> *"Unfortunately, group projects often suffer from "free riding" as economists call it or "social loafing" as it is termed by social psychologists".*
> (Maranto and Gresham 1998:1)

This occurs because some members of a group do not fulfil their responsibilities so that the others have to *"choose between working harder or accepting a poor project and a lower grade"* (Maranto and Gresham 1998:1). The Latané et al. social impact theory explains social loafing or hiding in the crowd, see section 3.9.7.2, (Latané et al. 1979). Another name for this is freeloading. (Boud et al. 1999; Amato and Amato 2005)

3.9.7.1 Free-riding

The free-rider concept "*was finally generalized analytically by Olson only in 1965 in his* Logic of Collective Action" (Maranto and Gresham 1998; Hardin 2003). Olson (1965) concluded that the free rider in a group received the same benefits as the others but without the *effort overhead*. He also noted, darkly, *"a surprising tendency for the exploitation of the great by the small"* (Olson 1965:35, original emphasis). Maiden and Perry's (2010) paper listed six *approaches* to dealing with free riders, while Amato and Amato (2005) claimed that free-riding it was *"the most frequently cited explanation for poor team performance"* (Amato and Amato 2005:42).

Maiden and Perry (2010) seem to have been a little naïve in using their 80/20 grading approach in dealing with non-contributors. It is also difficult to believe that this behaviour might not have been the urgent focus of an alternative GSA hypothesis. A literature search revealed nothing.

There also seems to be a contradiction in the Amato and Amato (2005) assertions. They noted that different styles were important for a balance of team attributes but they also noted that differing personalities were also potential areas of tension within the team.

3.9.7.2 Social loafing

Latané et al. (1979) coined a different term for the action, or more correctly the inaction, of non-contributing students. Social loafing occurs when individual effort decreases. This, they reported,

happens with increasing group size when individual effort is less easily identifiable. It also happens where the penalty for discovery is less than the cost of fully contributing. They suggested three causes of social loafing. First, they noted that task attribution and equity could be a problem where a team member attempts to maintain maximum effort for a disproportionate task allocated to them through a flawed task attribution system. Second, submaximal goal setting is the term they used where a group member re-interpreted the task instructions. This results in the group member deciding not to maximise their effort. They only do enough to appear to be complying with the instructions. Their compliance was only for cosmetic reasons. The third Latané et al. cause was lessened contingency between input and outcome. It meant that the social loafing team member believed that their lack of maximum effort had little bearing on their outcome, or module mark.

Both free riding and social loafing will result in the same student behaviour. For the purposes of this study, the terms are interchangeable. This is despite the Kerr and Bruun (1983) assertion that social loafing and free riding are clearly distinguishable concepts. According to them, social loafing occurs where *"group members exert less effort as the perceived dispensability of their efforts for group success increases"* (Kerr and Bruun 1983:78).

3.9.7.3 Strategic non-contribution

Another issue with GSA is the under-reported, deliberate strategic non-contribution by a group member.

Both Bacon et al. (1999) and Webb (1994) support the existence of this phenomenon. Bacon et al. reflected, *"Individuals may also feel that others on the team will do the work better than they will and so, feeling dispensable to the team, they reduce their effort"* (Bacon et al. 1999:472). As an aside, Bacon et al. also commented that *"In practice we have seen cases where students take turns loafing"*, (Bacon et al. 1998:70). This strategy of taking turns in non-contribution to the group effort would seem to be a quite sensible one, from a student point of view. It is a proper, democratic application of teamwork in practice, to ease the module workload. Similarly, Webb noted that:

> *"it is not clear whether students* [i.e. USA school-age pupils] *who did not actively participate in group work made a conscious decision not to interfere in the group's work,* ***for fear of slowing down group work or negatively impacting the group's score****".*

(Webb 1994:11 emphasis added)

Bacon et al. (1999) do not seem to have recognised that if some of the team members are most

able to do the project work, then the non-contributors to the technical content of the assessment product may contribute other equally important skills to the group. These could include monitoring and summarising progress. They could even be by housekeeping duties, managing the logistics of the group by arranging meeting times, venues and even the provision of equipment, refreshments, and calling for study breaks where appropriate.

In some circumstances, the other members of the group may even welcome non-contribution by low ability students. For example, this could be by implicit (or even explicit) agreement with the high ability group members. These more able group members may simply accept that doing the work themselves will give them a better chance of a higher mark and be more efficient.

3.10 Some views on GSA from students, staff, alumni and others

This section continues the GSA advantages and disadvantages theme from the previous section. It describes student, staff and alumni views on group assessment.

Morris and Hayes observed,

> *"As one would expect, both positive and negative comments were made regarding student experiences with group work. Some of the key points raised are summarised in the following comments …"*

(Morris and Hayes 1997)

The same study included a summary of academic staff views of group work. In general, they found it to be a useful learning and teaching technique. However, there they did find some dissenters. Some staff put forward the conjecture that it was overused. On the other hand, they also contended that groupwork in HE developed generic skills, encouraged deep learning, and improved students learning outcomes. They also agreed with Candy et al. (1994), see Figure 3, in section 3.5.2 above.

Leathwood et al. (1999) observed that groupworking was increasingly becoming an important part of UK HE learning and teaching. Morgan, in a later publication (2005a), highlighted that needing to develop students groupwork skills *"is not a new discovery"*. Bacon (2005) wrote more specifically of Business Schools. He commented that they *"often assign student group projects to enhance student learning of course content and to build teamwork skills"* (Bacon 2005:248). This is not, of course, a motive exclusive to business schools. It is also a coping response by universities. It is a response to increased pressure caused by rising student numbers and declining resources (Morris

and Hayes 1997; James et al. 2002). Some authors assert that group projects will help students to develop critical thinking, and interpersonal and communication skills, e.g. Candy et al. (1994). Knight was another author who suggested these, as well as other reasons for the popularity of GSA. He included

> *"increased amount of coursework, switch from summative to formative assessment, desire to develop and test a range of key skills, lack of time, and increased pressure on both instructors and students".*

(Knight 2004:64)

He also found that although students *preferred* individual assessment, they performed better at group assessment. The data from Knight (2004) is data set B in this study, see section 6.3.2.

In addition, Gibbs has observed, *"Group project work is one of the most common forms of student-centred learning. It is used for a wide variety of reasons"* (Gibbs 1995a:13). He also noted in another publication that he had found programmes where students *"beg to be allowed to work alone"* (Gibbs 1995b:6).

Barr et al. (2005) have also explored this loner phenomenon in the context of the classroom. They noted, *"The lone wolf appears to play a role in how teams function and perform"* (Barr et al. 2005:81). They also referred to teaming activities, rather than to group work. Their finding would also seem to be generalizable outside the classroom, indeed anywhere where any kind of group work is attempted.

Stephenson was the Thorley and Gregory (1994) book series editor. He observed that academics and employers hypothesized that collaboration could improve students learning experiences. It would enable the development of students' key skills by practical experience of group-dynamics, he noted. It also helps peers to learn to interact. As mentioned in section 3.9.1 however, he also asserted that groupwork *"can also be fun".*

Mello asserted that, at least in management education, *"The popularity of having students work in small groups can be traced to the fact that group work provides students with a number of benefits"* (Mello 1993:253). He did not elaborate on where this popularity occurred. It could have been with students, their 'instructors', the educational institution management, or any mix of these. He asserted that, from the student viewpoint, there were five main benefits of groupwork. First, it allowed them to gain practical experience of theoretical course material. The students would gain

much deeper insights into group-dynamics and processes. Second, using it could allow their 'instructor' to develop assignments that were more comprehensive. Third was that students had the chance to practise and develop their interpersonal skills. Fourth, it exposed students more closely to other points of view. This promoted group cohesiveness and synergy. It could also allow them an appreciation of different cognitive styles, personalities, motivations and, as he noted, 'perceptual processes'. A fifth benefit, Mellow asserted, was that it further prepared students for the real world. His ideas would seem to be far more generalizable than just to management education. The list of Mello's benefits of group work is summarized in Table 6.

Table 6: Benefits of groupwork from the student perspective (Mello 1993)

1	The students gain a practical insight into group-dynamics and processes
2	Group-work allows the instructor to develop more comprehensive assignments
3	The students can develop their interpersonal skills
4	The students are exposed to the viewpoints and behaviour of other group members
5	The students are further prepared for the real world

(Mello 1993)

Garland's (1996) staff responses to the question of benefits of small group learning were similar to those of the student perspective from Mello. Both studies were from a business management HE perspective. In her chapter on using research to improve student learning in small groups, she also noted that at that time, *"in higher education little evidence exists as to the effectiveness of student learning in small groups and to ways in which student learning may be improved"* (Garland 1996:224). Also, see Gibbs' (1995a) reasons for group work in Appendix 6.

Researchers of a 1997 Department for Education and Employment funded study, found concerns among alumni for groupwork assessment. The alumni found them to be *"exclusively based on outputs rather than contributions"* (Taylor and Rumpus 1997). The Taylor and Rumpus alumni suggested that GSA did not assess students on their reflections on the groupwork process. They found that students are more likely to be required to reflect on the process in an individually submitted assessment. The alumni comment that they would welcome more opportunity *"to be judged on their individual work"* is also informative. This suggested that more work is needed to integrate groupwork (and GSA) successfully into undergraduate study programmes. There was also *"a suggestion that students were perhaps required to undertake too much group work"* (Taylor and Rumpus 1997).

Arguably, the most graphic student description of their groupwork experience was "*I'd Rather Vomit Up a Live Hedgehog"* (Strauss 2001). This was the article title. The quotation dramatically

expressed the feelings of a level two student in a mainstream Australian university programme.

3.11 Degree classification research

In the research on degree classification, some authors, in an analysis of eight different degree subjects, have defined a good degree as a first or an upper second. They found "*a modest rise in first class degrees since the mid-1980s but a more rapid rise in the proportion of upper seconds"* (Myron-Wilson and Smith 1998:535). They suggested that some universities consistently award more good degrees than others do and proposed what they called a twofold *"tentative explanation"* for this increase. The first was that students (and presumably their advisors and other stakeholders) recognised that successive research assessment exercises (RAEs, to be replaced by REF, research excellence framework in 2014) were *"becoming both more valid,* [i.e. an important measure for stakeholders to consider] *and more widely known"* (Myron-Wilson and Smith 1998:538). They also noted that the overall increase could be due to advances in several areas of education. These included teaching quality and organisation. In turn, these resulted from better standards of textbooks, and of lecturer training. They also observed that students were also becoming more vocal and assertive consumers. The final improvement, they noted was the introduction of statutory quality assurance procedures. They also noted three other possibilities. First, that teaching syllabuses and examinations were co-ordinated and linked better than they had been previously. Second, modular structures may make exam questions more predictable. Third, *'semesterisation'* may help students to be more consistent in their pattern of study. These could of course just be somewhat optimistic viewpoints.

They used regression analysis with several of what they referred to as degree correlates. They found that the percentage of good degrees at each institution correlated with the:

Mean A-level score
Percentage of mature students
Staff-student ratio
Mean tutorial size
Amount of optional choice in programme
Structure
Extent of double internal marking
Extent of blind marking
Perceived direction of external examiner's advice
RAE ratings

Regression analysis treats the variables as being independent of each other. The authors found that *"The only significant correlate with the percentage of good degrees was the (1989) RAE rating (r = .45)"* (Myron-Wilson and Smith 1998:537).

They noted that assessment methods included the distinction between coursework and examination. A further conclusion from this present study supported that of Myron-Wilson and Smith that the choice of assessment method influences classification.

Assumption 17 (of 21) Oppenheim et al. (1967), which was mentioned earlier, included *"that examination results should be distributed in a certain way"* (Oppenheim et al. 1967:349). Also, see section 4.8. At that time, the term *examination result* was probably synonymous with *summative assessment result*. As part of assumption 17, they also assumed:

There will always be only very few Firsts.
The largest numbers will be from 2:1 and 2:2 degrees.
There will be very few failures.
There must always be some failures.

There were other parts to assumption 17. For example, they would regard as *'wrong'*, results where 40 percent of a subject cohort obtained Firsts. Finally, and tellingly, they noted that *"We also, in many subjects, make completely arbitrary assumptions about a pass mark"* (Oppenheim et al. 1967:349)

Professor van de Linde (2002), as vice-chancellor of Warwick University wrote, pre-Burgess, a spirited, brief and thought provoking article in *The Guardian*. It severely criticized the system of merely classifying degree passes into four categories (i.e. first, 2:1, 2:2 and 3rd). He offered a powerful and graphic sporting analogy to illustrate his point. In this, judges of a sporting contest simply told Jonathan Edwards (an Olympic gold medallist at Sydney 2000) that his triple jump was one of three categories. It was an *'OK jump, a good jump or a great jump'.* The judges did not disclose the actual distance jumped. He continued by reminding readers that the practice of using the four-degree classification system was not because of the lack of data on graduate accomplishments. It was from tradition, and the belief of many that the primary merit of the system was its uniformity across degree courses and universities. Professor van de Linde pointed out that

> *"Students should be given their actual marks, rather than a classification. Employers would be better informed and an unnecessary mystique would disappear. Every student would also strive, then, for the best possible marks - not just to scrape into a particular band".*

(Linde 2002)

It seems that first-degree stakeholders accept the meaning and integrity of a degree classification almost as an act of faith. Unfortunately, there seems to be little supporting evidence for this. It is similar to a belief in the monetary system. It only has a value because its stakeholders have faith

in its value. If, or when, that faith crumbles then the degree classification will become a devalued currency.

3.12 The Jumpstart groupwork option

For several years, White and Carr (2005c) have introduced students to groupwork in a novel, non-technical, non-summatively assessed manner. Their project was called *JumpStart* (JumpStart 2010). It was originally for undergraduates beginning their studies at Southampton School of Electronics and Computer Science, but was quickly extended to include all new students at Southampton. They are introduced to groupwork *by voluntary* involvement. In addition, this involvement is via an activity to allow newcomers to get to know their year group and the city of Southampton. It takes place during the introductory study week zero, *'Freshers'* week. For the year reported in the 2005 article for example, the event received excellent support from the students. In that year, 120 MA students who were predominantly from overseas joined around 300 new undergraduates for the *JumpStart* event. White reported that it was so successful that students formed their own students' union society, with only "*a little gentle prompting"* (White 2006). The scheme introduced students to the interpersonal skills and dynamics of groupwork while not directly affecting their own individual degree classification. As the *JumpStart* website acknowledged, its *City Challenge* aimed to get students learning and working more effectively (JumpStart 2010). They do this by putting them in an environment where, to gain the maximum benefit from the activity, they must collaborate in groups. This, the *JumpStart* team hoped, would also change the culture of *isolated strugglers,* which they observed, demanded so much teaching staff support.

3.13 Biographical predictors of student GSA outcome

Bacon et al. (1998) reported on biographical GSA group outcome predictors. From their quite small, single cohort, study, *"the average of the individual abilities on the team was found to predict student team performance"* (Bacon et al. 1998:63). They also found that team size had little effect, and gender diversity had no effect on team performance.

3.14 Andragogy

As mentioned earlier, in section 2.3.1, Smith (2008a) reported that Eduard Lindeman described his *" orientation as 'andragogical … - which appears to be the first English-language use of the term"*. The author also noted that the concept was almost two-hundred years old. It was *"originally formulated by a German teacher, Alexander Kapp, in 1833"* (Smith 2008b).

Some of the main points arising from the teaching of adults, which also seem appropriate to this study, are noted in *The Democratic Man: the selected writings of Eduard C. Lindeman*:

> *"None but the humble become good teachers of adults. In an adult class the student's experience counts for as much as the teacher's knowledge. Both are exchangeable at par. Indeed, in some of the best adult classes it is difficult to discover who is learning most, the teacher or the students. This two-way learning is also reflected in the management of adult education enterprises. Shared learning is duplicated by shared authority"*

(Gessner 1956:166)

Also as mentioned earlier, in the mid 1990's McKenna published a short series of articles on learning theories as they applied to nursing training. She noted that Knowles *"developed the theory of andragogy, the art and science of teaching adults"* (McKenna 1995:31). She also noted that he also asserted that pedagogical teaching methods were inappropriate for adult education. His basis for this was similar to Lindeman's view. It was with reference to the importance of the wide range of experiences that as adults, students bring to their learning. More importantly, McKenna also cited four main differences of andragogical learning. First, adults need to be able to apply what they have learned. Second, they have a wealth of personal and life experiences to draw on as part of their education. Third, their learning involves an investment of self so any new learning will affect that self-concept. Fourth, they are strongly self-directed so their education should accommodate this.

3.15 Chapter summary

This chapter has reviewed some of the background literature linked to education.

In section 3.1 the terms *group* and *team* were defined as they relate to this present study. It was also noted that there seems to be little commonality of definition among authors. For this present study, as well as many others, the terms groupwork and teamwork were interchangeable concepts. As are group work and team work. In some situations, the semantics may be important, but it is not the focus of this thesis. The favoured HE term when applied to students is group. There was also an acknowledgement that this concept of a team was not new; and that while a team must be a group, it is clear that a group does not have to be a team. Section 3.3 was on the practise effect for GSA students. Section 3.4 explained the ubiquitous nature of GSA in higher education. That being so, then GSA must surely be included.

Section 3.5 outlined the five main reasons that make up the rationale for GSA practice. The first

was that it is what employers want, although this seems not to be supported by all employers. The HEA reported a *'broad consensus'* among employers, while the RSA appeared to be reporting the opposite of this. In addition, Wolf's observations indicate the complexity of the issue. The second was that it teaches generic group working skills. Third was that it is an effective learning and teaching tool. Fourth, it allows more meaningful and realistic projects. Fifth, resources may be used more efficiently. Section 3.6 reported on forming student module groups and allocating group members. It included three methods of group membership assignment, and illustrations of group sizes and learning styles. Included in section 3.6.3.3, were findings on GSA heterogeneous versus homogeneous personality groups. Section 3.7, reported on the impact of GSA on students' marks from the literature.

Section 3.8 was the unverified claim that GSA tended not to disadvantage high achievers, which, if verifiable, would have supported part of the findings of this study. Section 3.9 was about the main disadvantages of GSA; fundamentally, groups do not receive collective degrees classifications. Individuals receive individual ones. It included the problematic nature of GSA, Adam smith's 'invisible hand', a legal issue arising out of GSA practice, and the nature of *groupthink*. The Tuckman forming, storming, norming and performing model was included. As was the non-contribution to the group effort, e.g. the free rider or social-loafing effect, and tactical non-contribution, and the GSA higher mean marks and lower standard deviation finding.

Section 3.10 illustrated the mixed views of student, staff, alumni and others on GSA. Conclusions from the Gibbs, Mello and Garland studies seem to be in general agreement. This is despite, inevitably, the contexts of their studies being slightly different. Section 3.11 was on the rather limited degree classification research. Section 3.12 outlined the Southampton University *Jumpstart* project initiated by Carr and White. It introduced new undergraduates to groupwork in a manner that is non-contributing to their future degree classification, and continues to do so. Section 3.13 was from the limited literature reporting biographical predictors of student GSA outcome. It reported that student team performance was predicted by the average of the individual abilities on the team. Section 3.14 presented observations on the definition of, and stakeholder understanding of, the term andragogy.

The general background literature appears to have concentrated on group and peer learning and assessment, how groups are formed, learning styles and rationalising GSA for pragmatic reasons.

There is, as mentioned elsewhere, almost no literature directly investigating the impact of groupwork on the overall student mark, and therefore its impact on their award classification.

The next chapter reviews the literature on summative assessment in HE. The purposes of HE and the purposes and roles of assessment are reviewed. There are sections on the role and definition of summative assessment and GSA use. The validity and reliability of HE assessment is also reviewed. A short section on the use of Portfolios is also included. The assumptions of Oppenheim et al. underlying the use of university examinations, the *Standards* (Novick 1985) fairness chapter and the importance of assessment in higher education end the second and last literature review chapter.

Chapter 4. Review of higher education summative assessment literature

The previous chapter was a review of the literature on general issues of education that provided the broad background to this study. This chapter reviews issues specific to HE summative assessment.

4.1 Chapter introduction

Records of education assessment seem to have begun at the time of Confucius, between 551 and 479 BCE (Huanyin 1993). At that time, education assessment was carried out in order to warrant individuals for posts in the Chinese Civil Service. As Huanyin noted, *"This approach also helped to create the conditions whereby the emergent land-owner class could accede to the authority conferred by learning and produce talented men* [SIC] *from its midst"* (Huanyin 1993:212). More recently, Delandshere viewed formal assessment as a *'socio-political practice'* and observed that *"procedures have existed for centuries"* (Delandshere 2001:116).

Several authors have expressed their views on the importance of contemporary educational assessment, for example Novick (1985), Gibbs (1994b) and Brown et al. (1996). It has also been observed that students could not escape the effects of poor assessment (Boud 1995b). This Boud quotation was cited, and quoted in full earlier in the introduction to Chapter 2. The author has echoed this point in other papers, e.g. Boud (1998). Even when Dearing (1997), misquoted the original by omitting the clause *"(by definition if they want to graduate)",* it still seems to hold its potency.

Barfield (2003) studied group assessment in particular rather than educational assessment in general in the USA. His data subjects were *"A cohort of 230 students from a large southern metropolitan university enrolled in sections of Group Interaction and Decision Making and Conflict Management classes"* (Barfield 2003:355). In addition, his *"college classrooms and the workgroups within them … comprised of students of wide age ranges, different responsibilities and obligations and diverse cultural, racial, ethnic and socio-economic backgrounds"* (Barfield 2003:356). His study was focused on students' perceptions of and satisfaction with what he called group grades, and on the group experience in the college classroom. He found three student characteristics that were related to group grade:

The less group grade experience that a student had the more likely they were to agree that everyone in the group deserves the same group grade.

Students who work part-time were more likely to think that a group grade is a fair assessment of their contributions than students who work full-time.

Older students were more likely to be dissatisfied with a group grade experience than middle and younger age students.
(Barfield 2003:355)

Orr (2010) studied students' experiences of group work assessment in the creative arts. She explained, *"Group work is central to pedagogy in the performing arts"* (Orr 2010:304). Following on from Delandshere's view (given above at the start of this section), Orr also explained, *"assessment is a socially situated practice informed by, and mediated through, the socio-political context within which it occurs"* (Orr 2010:304).

Other authors who wrote about academic achievement in general rather than specifically about groupwork or GSA included Entwistle and Wilson. They made the point in 1977 that:

> *"it was already clear that we would not find any single variable which could explain more than a small proportion of the variations in academic performance. Many factors are involved and it is only from an examination of how these interact that progress can be expected".*

(Entwistle and Wilson 1977:5)

As part of keeping up to date with current research in the topic during the present study, continued on-line literature searches were made. The searches produced very disappointing results. These often returned the screen message: *Your search … did not match any documents*. The matches that were found were for psychology papers and concerned patient therapy groups rather than student group work in higher education.

There are nine further sections and a summary section in the remainder of the chapter. Section 4.2 outlines and reviews the literature on the purpose of higher education. Section 4.3 concerns the roles and purposes of assessment in higher education. It includes Newton's warnings of there not being one assessment method suitable for all applications. Section 4.4 is on the role and definition of summative assessment. Section 4.5 covers GSA use. Section 4.6 focuses on the validity and reliability of HE assessment. Section 4.7, reviews the use of portfolios. Section 4.8 describes the Oppenheim et al. (1967) *Assumptions* underlying the use of university examinations that were relevant to this study. Section 4.9 concerns the joint American Education Research Association (AERA), American Psychological Association (APA), and the National Council on Measurement in Education (NCME) educational and psychological *Standards* (Novick 1985) views on fairness. Section 4.10 illustrates the importance of assessment in higher education. There is

also a summary section (4.11).

4.2 Purposes of higher education

It has been observed that *"Few individuals would deny that learning is the primary purpose of higher education and that teaching is the foremost means by which that goal is accomplished"* (Tait 2009:192). It has also been suggested that there were four purposes of higher education (Atkins 1995). The first was that it gave a general education experience that had an *"intrinsic value in its own right"* (Atkins 1995:25-26). Second was to *"prepare students for knowledge creation, application and dissemination".* The third purpose that Atkins purposed was to train students for a particular profession or vocation. Her fourth purpose was to prepare students to take their place in the workforce in general. Additionally, she questioned what it was about a degree that was distinctive.

Like Atkins (1995), the Dearing (1997) report also noted four main purposes of higher education. These were more abstract and broader ranging. The first was *"To inspire and enable"* (Dearing 1997:5.11). This was to maximize students' capabilities throughout their lives. They could then, the report claimed, grow intellectually, be well equipped for work, could contribute usefully in society and *'achieve personal fulfilment'.* Another purpose was citizenship based which was concerned with future societal needs. Dearing also included learning for its own sake and for the economic benefit of society as a purpose of higher education. From these it seems clear that in the opinion of the Dearing (1997) committee at least, summative assessment is merely peripheral to the main purposes of higher education.

It has also been noted that *"assessment defines what students regard as important, … It follows then that it is not the curriculum which shapes assessment, but assessment that shapes the curriculum and embodies the purposes of higher education"* (Brown and Knight 1994:12). The first part of this is also fundamental to the next section, which reviews the roles and purposes of assessment in higher education.

4.3 Roles and purposes of assessment in higher education

It has been suggested that assessment had four separate roles.

> *"formative, to provide support for future learning; summative, to provide information about performance at the end of a course; certification, selecting by means of qualification; and evaluative, a means by which stakeholders can judge the effectiveness of the system as a whole".*

(Hornby 2003b:3)

In another paper the same year, Hornby noted that these different roles could create tensions for universities. For example, *"awarding more first class degrees may indicate rising standards of teaching and learning or* ***precisely the opposite****"* (Hornby 2003a:435, emphasis added). (There is also a review of the literature on degree classification research, including the issue of the rising numbers of good degrees, i.e. degree inflation, in section 3.11.) In addition, student marks for more challenging, or to use Hornby's term *hard* topics, tended to use the whole scale range. These were topics such as mathematical science. His assertion was that confident markers would also use the whole scale range. *Soft* topics, or those marked by those who want to avoid the extremes of the marks range, could be marked towards the mean. Hornby however, found no evidence of stakeholders views influencing this. He called it risk averse/defensive marking behaviour. He concluded that the need to satisfy stakeholders has led to *"Increased transparency"* and that this is "*revealed as one of the most important influences on marks/grade"* (Hornby 2003a:451). He also noted, that out of a sample of almost 400 business studies students *"The data for these degree students shows ... for 14 out of the 22 final year modules in these degrees, over 50% of the percentage mark scale is not used"* (Hornby 2003a:451).

With Hornby's four assessment roles (Hornby 2003a), his view of the range of stakeholders of first-degrees does seem to have been rather limited. He only included colleagues, external examiners and students. He did not seem to consider that, for example, students' families, schools or potential (or actual) employers or society in general could be stakeholders (see section 2.1).

The anonymous contributor to the JISC (Joint Information Systems Committee) InfoNet web page *'What Do We Mean by Assessment',* (JISCinfoNet 2007), suggested that there two main *purposes* of assessment within both FE and HE. The first was to assist learning. They noted that this was done by trying to make the assessment relevant to the overall goals of the module and by making it part of the learning process. The second purpose was to evaluate the education system in terms of its effectiveness. The author observed, *"we must be able to determine not only the overall learning but which areas are not effective and need modification"*. They also suggested ten reasons that tutors had for assessing students. Nine were student centred. Only one concerned evaluating the programme or module. This is shown as reason 10 in Table 7.

Table 7: Reasons tutors assess students (JISCinfoNet 2007)

1	To pass or fail a student
2	To grade or rank a student
3	To select for future courses
4	To predict success in future courses
5	To provide a profile of what a student has learnt
6	To diagnose students' strengths and weaknesses
7	To provide feedback to students to improve their learning
8	To help students to develop their skills of self-assessment
9	To motivate students to provide feedback to teachers
10	To evaluate a course's strengths and weaknesses

Purposes of assessment have also been referred to as *'assessment judgements'* (Newton 2007). He noted that *"There is no summative purpose … There are no formative judgements"* (Newton 2007:156-157). He also listed a broad selection of uses for assessment judgements. Adding that *"Whereas, in the past, we have tended to want to classify them into a smaller number of categories, it is probably more constructive to consider each a category in its own right"* (Newton 2007:149). He developed eighteen categories of uses for assessment *'judgements'*. These are shown in Table 8.

Table 8: Educational assessment uses categories (Newton 2007:149)

1	Social evaluation	10	Licensing
2	Formative	11	School choice
3	Student monitoring	12	Institution monitoring
4	Transfer	13	Resource allocation
5	Placement	14	Organizational intervention
6	Diagnosis	15	Programme evaluation
7	Guidance	16	System monitoring
8	Qualification	17	Comparability
9	Selection	18	National accounting

Newton's (2007) article focused on school age children. It was aimed primarily at educational policy-makers. This does not detract from the import of his article to this study. He introduced it with a warning to bear in mind the purposes of the systems. *"The fact that a system which is fit for one purpose will not necessarily be fit for all purposes is a fundamental consideration when evaluating the legitimacy of proposals"* (Newton 2007:149). This was one of the recurring themes in his paper. Fitness for one purpose did not imply fitness for any others. He also thought that even the phrase *assessment purposes* had at least three interpretations, i.e. level of judgment; decision level; and impact level. The outcomes need not necessarily be as intended, or even as anticipated, by the designers of the assessment systems. Newton noted that Scriven first used the terms summative and formative to highlight two different methods of programme evaluation, (see section 4.5.1). In addition, he also issued an appeal to assessment system designers that they should try to ensure that results were not used improperly. They should *"identify, for all stakeholders, those purposes for which results are unfit"* (Newton 2007).

Newton's plea for researchers to identify for stakeholders, the '*purposes for which the results are unfit',* was well articulated (Newton 2007) and seems relevant to GSA. It remains to be seen whether his plea will be heeded. A similar plea seems to have been largely unsuccessful for Oppenheim et al. (see section 4.8).

These views are clearly supported by AERA, APA and the NCME in their *Standards* (Baker et al. 1999). These standards do however use the terms, tests and testing, rather than assessment. They define assessment as *"Any systematic method of obtaining information from tests and other sources, used to draw inferences about characteristics of people, objects or programs"* (Baker et al. 1999:172). They require that tests and testing programmes be developed on a sound scientific basis (Baker et al. 1999:43, standard 3.1). Their observations about fairness in testing are a separate topic that is reviewed in section 4.9.

It has also been observed (LondonMet 2006: section A) that assessment should support student learning and be fit for purpose. The same authority also asserted that it should be valid; reliable; feasible; efficient; adhere to professional standards and be fair and transparent.

It has also been noted that assessment procedures offer answers to the questions:

> *"What student qualities and achievements are actively valued and rewarded by the system? How are its purposes and intentions realized? To what extent are the hopes and ideals, aims and objectives professed by the system ever truly perceived, valued and striven for by those who make their way within it?"*

(Rowntree 1987:1)

4.4 What is summative assessment?

According to one at least professional institution, i.e. the Chartered Institute of Educational Assessors (CIEA), summative assessment is *"also known as* 'assessment of learning' *or* 'assessment of attainment'*"* (CIEA 2009). This was in comparison to formative assessment. They also observed that formative assessment *"may also be referred to as* 'assessment for learning' *or* 'assessment for teaching'*"* (CIEA 2009).

Summative assessment has also been called sustainable assessment (Boud 2000). The author noted that it *"can be defined as assessment that meets the needs of the present and prepares students to meet their own future learning needs"* (Boud 2000:1). Another of his observations was the ways in which assessment has to do what he called *double duty*, which confounds Newton's

point cited in the previous section. Boud noted that they have to encompass both formative and summative assessment. The former was for learning and the latter was for certification. They also have to *"focus on the immediate task and on implications for equipping students for lifelong learning in an unknown future"* (Boud 2000:8). These observations were concerning learning to learn, as well as being about assessing the student's knowledge of the subject discipline. To Boud, it also involved identifying appropriate standards and criteria and, against these, making judgments about the quality of students work. In addition, assessments have to address both the learning process and the content domain. At the same time, he warned that assessment *"always performs functions other than the ones teachers and examiners normally think about and take account of"* (Boud 2000:8). (Also, see observations from Mutch and Brown (2001) in Appendix 7.)

It has also been observed that summative assessment also has a feedout function (Knight 2002). Knight noted that stakeholders use it as a performance indicator and predictor. These stakeholders included students, departments and higher education institutions. They also included employers, funding bodies, quality agencies and compilers of league tables (see Appendix 1). He also noted that *"So important are those feedout functions that such assessment is often called high stakes or summative"* (Knight 2002:276). In a later publication (2008), he and Yorke noted how high-stakes summative assessment resulted in *local meaning*. They seem to have meant that there was no agreed global definition of it. They saw it as a short summary of achievement. It indicated, for example, potential future performance. It may also attest to fitness to practice. They also pointed out that *"Warrants may be short but they are not simple compressions: it is not possible to* 'double-click' *on the warrant and* 'unzip' *full information about achievement"* (Knight and Yorke 2008:76). The textual content of the *'warrant'* may vary. It may be a very brief degree class number. Alternatively, it may also include a detailed transcript.

A more straight-forward and elegant meaning of summative assessment was defined by Brennan (2008). He observed that for students, it *"means any assessment that generates a mark which directly contributes to a result on a university transcript"* (Brennan 2008:53). This was the year before the HEAR was introduced in the UK (Paton 2009).

Another author made an equally telling observation on the role of summative assessment. He believed that it was *'crucial'*, and that *" We get the kind of learning that our assessments deserve, since students will decide what is appropriate"* (Stobart 2008:87). This further supports the Strauss

(2001) appeal for caution when using GSA techniques, (see section 3.9.1).

4.4.1 Criterion referencing and norm referencing

Two types of summative assessment recognised in the literature are criterion referenced and norm referenced. Criterion referencing measures *"The degree to which … achievement resembles desired performance at any specified level"* (Glaser 1963:519). On the other hand, norm referencing refers to *"the capability of a student compared with the capability of other students"* (Glaser 1963:520). Glaser noted that *"Educational achievement examinations, for example, are administered frequently for the purpose of ordering students in a class or school, rather than for assessing their attainment of specified curriculum objectives"* (Glaser 1963:520).

4.5 Assessment and GSA practice

This section on the use of GSA is in six parts. Section 4.5.1 reviews the dual purposes of assessment as they have been presented in the literature. Section 4.5.2 considers two methods of summative assessment, peer- and self-. Section 4.5.3 relates to alternative methods of assessment. Section 4.5.4 looks at one author's view (David Boud) of educational assessment. Section 4.5.5 focuses on methods of deriving individual marks from group project marks. Section 4.5.6 concerns the place that summative assessment holds as a student motivator.

4.5.1 Assessment purposes: summative and formative

(Also, see for example Appendix 7, table of Purposes of Assessment (Mutch and Brown 2001:5).)

As noted in an earlier section, the first author to use the terms summative and formative in regard to student assessment seems to have been Michael Scriven (1973:63-64). He used the word evaluation as a synonym for assessment. He tells us that his "first publication in evaluation introduced the term *formative evaluation* and *summative evaluation*" (Scriven 2004). That he was the first is also cited by for example Boud (1995b), Taras (2005) and Newton (2007).

He took the view that *evaluation "is a peculiarly self-referent subject"* (Scriven 1983:229-230). Part of his concept was that it is like the sociology of science. For him, the sociology of science included the sociology of the sociology of science. Hence, he observed, it is self-referent. Evaluation, he asserted, applies to the process and artefacts of all serious human endeavour and this must include evaluation. These were originally terms for the assessment of the programmes themselves. They were not about students' academic progress within them.

Scriven also wrote of his doubts about his own definitions. He observed that a previous note of his

that all evaluations that were not formative, were summative, was too rigid. Sometimes in the assessment of programmes, he asserted, the assessor would not be *'connected'* to the process. As an example of this he explained that *"a historian may evaluate the use of the draft by the U.S. government in the Vietnam War without any intention of advising anyone on how to improve it or advising anyone on whether to fund it, defund it or indeed to resuscitate it"* (Scriven 2004:187-188). He also expressed a readiness to use another term. He suggested *"*"ascriptive evaluation" *until I think of a better one".* From this, his thinking had clearly moved on between 1983 and 2004. As Alkin (2004b) reminded readers in the introduction to the same book

> *"theorists' views are not fixed in time as would seem to be implied by reading their published works. Theorists typically change their views over time, and published work often lags behind these changes."*

(Alkin 2004b:7)

4.5.2 Summative assessment types: Peer- and self-assessment

Quite a number of authors have attested to both peer- and self-assessment often being included in GSA marking schemes (see Appendix 5). Falchikov's (1986) article, for example, was based on the results from just such an experiment. She wrote about *Product Comparisons and Process Benefits of Collaborative Peer Group and Self Assessments*. She observed that *"Staff exercise unilateral intellectual authority"* (Falchikov 1986:146). For her, *"Any student assessment procedure should meet a number of criteria. It should be - valid, reliable, practicable and fair, and useful to students"* (Falchikov 1986:146). She noted that by excluding students from each step in the decision making process, the power relationship was all in favour of the assessor. She also observed that all the power was with the staff. They decided what students should learn and they designed the programme of learning. They also determined the assessment criteria, and they assessed the students. Falchikov also noted that this system was unpopular with many tutors and students. In addition, for her, *traditional* methods also had potential negative effects. She noted that *"in spite of its widespread use, the system is notoriously unreliable"* (Falchikov 1986:146). She argued that this gave rise to a number of injustices. In her view, it led to poor inter- and intra-rater correlation of assessment scores. In turn, she observed that students were preoccupied with their assessment marks. This meant that they excluded any work activity not relating directly to them. It hindered the growth of both their responsibility and their autonomy. From her point of view, it led to conformity and put off personal development. It also, she asserted, discouraged them from developing their key skills. Falchikov observed that the traditional system generated an inappropriate study motivation among students. It emphasised external rewards and punishments. She explained that this motivation was extrinsic. It was also, she noted, *"intellectual alienation"*

(Falchikov 1986:146).

Heron (1988) developed the Falchikov disempowerment argument further. He took the view that use of authority was always contrary to the aims of education. It was often also self-perpetuating. He argued that, *"Staff unilaterally assess students, some of whom go on to become staff and unilaterally assess more students. … Unilateral control and assessment of students by staff means that the process of education is at odds with the objectives of that process"* (Heron 1988:79).

There was no further discussion on this by either author. For example, they did not elaborate on how every student could know what subject topics they should learn, and how they might have gained sufficient experience of programme design to be able to design their own programme of learning. Despite this, their concerns remain relevant to this present study. Heron noted that outputs from the university education process included educated people. Because of this, among other things, they would be self-determining and would then be able to set their own learning objectives and plan how to arrive at them. They would be able to decide how to measure these learning objectives. They would also be able to assess themselves *"in the light of these criteria"* (Heron 1988:79-80). Graduates should have the attributes of good academics, he asserted.

He also argued for other assessment methods. Traditional education fell short of his aim of student self-assessment. He noted that in each learning objective, the staff did it either for or to the students. Academic staff also had unilateral control of students' assessments. Their aim was to produce other genuinely self-determining academics. He argued that this aim was incompatible with the methods of summative assessment. (Heron 1988:79-80)

Like Falchikov (1986) and Heron (1988), Gale et al. (2002) also had issues with the focus of power when students were summatively assessed. Theirs was a triadic study. It used self-, peer- and tutor-assessment. They also closely examined *"Questions of power and knowledge"* (Gale et al. 2002:557). They quoted Brew's (1999) book section at some length. It concerned using self and peer-assessment to lead to autonomous student assessment. In particular, Gale et al. noted,

> *"We know that assessment enables the exercising of power by tutors over students. … In this context, scepticism about traditional power relationships points to the need to involve students in sharing power by assessing their own and each other's work".*

(Gale et al. 2002:557-558)

The paper highlighted the problematic nature of self- and peer-assessment.

Underlining this problematic nature, Sluijsmans and Prins (2006) warned, *"It should be noted that peer assessment skills are not easily and automatically acquired. Peer assessment is considered a complex skill that needs to be developed"* (Sluijsmans and Prins 2006:9). They also warned that neophyte peer assessors may be unsure, and need more help in learning the necessary skill. Their note was also cited in Gielen et al. (2011:149).

As Gibbs put it, in an on-line contribution to the debate:

> *"It is the level of challenge and standard of the work of individuals which normally determines their marks rather than their effort. The issue to be addressed by the teacher is how to distinguish marks for individuals within a group who have made different kinds of contributions."*
> (Gibbs 2009: 5)

What he meant by the level of challenge is unclear from his article, but what is clear is that, from students, he supported the importance of the quality of effort, i.e. the *'standard of the work'* over that of the quantity of work. (Also see discussion in section 8.2.6.)

4.5.3 Alternative assessment methods

It has been noted that the *"repertoire of assessment methods in use in Higher Education has expanded considerably in recent years"* (Sambell et al. 1997:351). They cited common terms for alternative assessment. These included:

> *"performance assessment, authentic assessment, direct assessment, constructive assessment, incidental assessment, informal assessment, balanced assessment, curriculum-embedded assessment, curriculum-based assessment".*
> *(Sambell et al. 1997:351)*

Use of a flexible method of HE assessment has also been reported, see Cook (2001). Cook's students had two optional assessments and only the end of semester exam was compulsory. The marks used were the highest from any of the resulting four combinations of marks for the semester. This allowed her students to choose to do one or both of the optional programme assessments, or neither of them. She claimed that this offered the students more choice. Other alternative forms of assessment listed include projects and investigations, varied writing assignments, oral assessment, realistic or problem-solving tasks, simulations, portfolios, profiles, group assignments, self, peer and co-assessment, and Wikis and podcasts. (Race 1995a; Sambell et al. 1997; Barrett 2004)

4.5.4 Boud on educational assessment

In 1990 David Boud warned that *"many current assessment practices are incompatible with the goals of independence, thoughtfulness and critical analysis to which most academics would subscribe … we certainly cannot be sanguine about our assessment practices"* (Boud 1990:101).

His article discussed assessment practices in general. It was not specifically on the topic of GSA. Even so, the point he made was important to this present study. He also noted that popular assessment methods were not consistent with those that academics applied to their own studies. He noted that many departments had an assessment policy that undermined deep learning. This might, he asserted, have unintended consequences in assessment. It might lead to short-term surface learning in a pragmatic attempt to maximize marks.

In a later article (Boud 1995b), he suggested that students would vary their learning styles in response to the assessment feedback cues and messages they received from their teachers. As he saw it, despite staff good intentions, assignments would often lead to a narrow, instrumental approach to learning. It would emphasise the reproduction of the lecturers' thoughts as regards what was important in the syllabus. He also observed that it would discourage the hoped-for student critical thinking. It would do the same to deep understanding and independent activity. It would also discourage them from taking initiatives. Instead, they would take a pragmatic approach and follow their teacher's interpretation of the syllabus. Boud had earlier explained, *"They spend their time* 'swotting for examinations' *rather than trying to internalize and make sense of the subject"* (Boud 1990:104). The title of a section of a different Boud article, *"Assessment always leads to learning. But intended or not?"* (Boud 1995b:36) made the point slightly differently. In the same way, *"Every act of assessment gives a message to students about what they should be learning and how they should go about it"* (Boud 1995b:36). He also asserted that *"The message is coded, is not easily understood and often it is read differently and with different emphases by staff and by students"* (Boud 1995b:37). He followed this by observing that. *"The message is always interpreted in context and the cues which the context provides offer as much or more clues to students than the intentions of staff, which are rarely explicit"* (Boud 1995b:37). He also noted that students are not simply responding to the specific module. They were also remembering the lessons of their experiences of learning and assessment far beyond their university study. One of his contributions to the importance of assessment argument was the point that

> *"Assessment is the most significant prompt for learning. One of the most important outcomes of research on student learning is the recognition that learning must fundamentally be seen as relational. That is, learning is a function of both teaching and the context in which it occurs".*
>
> (Boud 1995b:36)

The Snyder (1971) *Hidden Curriculum* supported Boud's comments on student pragmatism, and of their surface approaches to study. He also contributed to the assessment versus learning debate.

In his *book*, he noted that in some circumstances with some modules some students would learn to be pragmatic and adopt surface approaches to study. In others, they may adopt deep or strategic approaches. He noted that the type of assessment and the nature of the tasks would help them to decide a strategy. They may learn that in many circumstances, in order to maximise their marks, they should use rote learning (Snyder 1971:32).

4.5.5 Methods of deriving individual marks from GSA

There is an extensive literature on methods of deriving individual marks from group marks. (See bibliography in Appendix 5.) There may also be as many variations in the methods of allocating group project marks, as there are convenors of GSA modules. For example, Brown et al. (1997a:136-137) have compiled a list of suggestions for group marking schemes, see Table 9.

Table 9: Suggestions for group marking schemes (Brown et al. 1997a, figure 8.11)

1	Everybody gets the same mark.
2	The group decides the marks of individuals at the end of the project.
3	The group decides the marking criteria at the beginning of the project and allocates marks at the end of the project.
4	The group allocates roles and marking criteria for each role at the beginning of the project. It allocates marks at the end of the project.
5	The group agrees that everybody will contribute equally to each task of the project. At the end of the project, those that did little get marks below the mean, those that did more get marks above the mean. The decisions are taken by the group.
6	The tutor plus the group use any one of schemes 2-5.
7	Only the tutor marks using one of schemes 2-5.
8	Individual viva.
9	Individual project exam mark plus group project mark.
10	Yellow card and Red card method. The group gets the same mark. If the group reports a malingerer, he/she gets a Yellow card and his/her mark is reduced by 10 percent. If there is no improvement, a Red card is issued at the end of the project and the student gets zero.
11	Everybody gets the same mark for the project but additional marks are given for individual contributions to it.

(Brown et al. also note that variations on these schemes are possible.)

For degree programmes put forward for accreditation by the British Computer Society (BCS), group projects must allocate individual marks. In their guidelines, they acknowledge what they called *"major activity being undertaken as a **group** enterprise"* (BCS 2007:14, emphasis added). In such cases, they require *"that the assessment is such that the individual contribution of each student is measured against the learning outcomes",* (BCS 2007). It is reasonable to suppose that they are not alone in this requirement, for an HE degree programme module seeking accreditation by a professional association.

Also from the UK, Michaelson (2004) has suggested five ways to allocate group projects marks to students. One, staff assess the group work. Two, the group assess their own project. Three,

individuals assess themselves and their contribution to the group. Four, other members of the group assess each individual's contribution to the project. Five, other groups assess the group. Michaelson also noted that methods 2, 3 and 4 could also be combinations of both self- and peer-assessment. In addition, she noted that assigning individual marks was the main issue with GSA *"when the group work forms part of the overall assessment of the individual"* (Michaelson 2004:2). In addition, in Burd et al. (2003) the authors explained that their paper *"Specifically concentrates on the development of an **effort** weighting from which individual marks are composed"* (emphasis added), rather than developing a marking scheme algorithm based on both effort and ability.

4.5.6 Summative assessment: Study motivator and primary study strategy focus

Many authors have attested that assessments have always been crucial to students (e.g. Brown and Knight (1994), Ramsden (1992) and Newstead (2002:70)). It is seen as both a strong motivating factor and a primary focus in their study strategy. Rogers for example, asserted that student *"attention* [to their academic studies] *is contaminated by the two questions, What are this teacher's leanings and biases towards this subject, so I can take the same view in my papers … What is she saying that it is likely she will ask on the exam"* (Rogers 1983:24). Another author noted simply that the requirements of the assessment are often the primary focus of the student (Ramsden 1992). One of Gibbs' many contributions to the student group work and GSA debate was his assertion that *"assessment is the most powerful lever teachers have to influence the way students respond to courses and to behave as learners"* (Gibbs 1999:41). Newstead (2002) also confirmed, in his abstract *"assessment has a major influence on students' attitudes towards their studies. Indeed, students' principal motivation is probably to get good marks rather than to learn about their discipline"* (Newstead 2002:70).

The view of Knight (1995) was that *"Assessment is a moral activity".* He also underlined this by noting that what and how we assess sends a message to students about what is valued. Additionally, he posited that assessment was the basic building block that supported their learning, *" the way students are assessed is the 'DNA evidence' of their learning experience"* (Knight 1995:13). In a later publication (Knight 2000:237), he argued the case for a programme-wide approach to assessment. Additionally, his plea for graduate transcripts predated HEARs (Paton 2009). He observed that, *"If degree awards are to have high exchange and use value then the assessments on which they are based need to be valid".* He also pointed out that this means, *"they should describe attainments in terms of both key, or transferable, skills and of subject-specific*

achievements" (Knight 2000:237). Knight's (1995) assertion of assessment as a *'moral activity'* is a reminder that, for summative assessment, there may be no absolute right or wrong responses, except perhaps in mathematics. Additionally, his use of the term *'DNA evidence of their learning'* seemed to be confirming assessment as the basic building block of student learning.

It has been noted that because summative assessment is such an important student study motivator and the primary focus of their study strategy, the art of assessing *"needs to embrace several different kinds of activity"* (Race 1995a). Race pointed out that, of course, assessment can have many forms. In addition he suggested that it could be argued that the more diverse the assessment methods, the fairer the, presumably overall, assessment is to students.

It has also been pointed out *"That assessment is central to student performance in higher education is a given"* (Ritter 2000:308). Like Boud's double duty (see section 2.6), she suggested that it has a dual function. Unlike him however, Ritter considered this problematic. The formative role was, for her, in facilitating, supervising and monitoring learning. The summative role was in *credentialing* graduates. This required giving the group assessment deliverables, the products of their abilities, an aggregated score for the general use of all stakeholders.

Credentialing entails at least one assessment task measuring student competence. This is usually administered at or towards the end of a module study period. It is also set against specified learning outcomes. Traditionally, Ritter (2000) noted, this was through *examinations*. Citing Brown & Knight (1994), she suggested that this was because they are *"guaranteed to test the individual on his or her own"* (Ritter 2000:308). Again citing Brown & Knight, she also asserted, *"These exams are "cost effective, easy to administer and usually thought to be objective and fair"* (Brown and Knight 1994; Ritter 2000:308).

Horowitz (1986) reported on the history of college student cultures from the late 18th century up to the 1980's. Her sample was of students in the United States. She noted the protests of the late 1960's. As soon as they were over, she asserted, grades and grade point averages returned to dominate student life even more than before. In writing of the early nineteen-eighties, she observed, *"grades* [i.e. summative assessment marks] *are the ultimate value"* (Horowitz 1986:35).

Horowitz' point about the return to dominance of grades in student life supports that of Becker

(1995, 1968). In Becker's 1995 essay he confirmed that *"Though historically contingent, taking somewhat different forms from time to time, the emphasis on grades has persisted"* (Becker 1995). He had collaborated on a book that included a chapter devoted to defining organizational rules and the importance of grades. The authors (Becker et al. 1968) discussed how grades were, as they put it, the currency of the campus. They stated that they were the *"chief institutionalized valuable of the college"* (Becker et al. 1968:55). They noted that grades were the only *'valuable things'* that students could acquire from the formal system of the HEI that applied to every student. Earlier in the book, they had been careful to explain where they had taken their sample. *"We studied the perspective students develop towards academic work in a particular place – the University of Kansas"* (Becker et al. 1968:15).

Haywood's (2000) contribution was the note that *"It is now well understood, as it was not before, that assessment has a powerful influence on student learning"* (Haywood 2000:9). He expanded this influence of assessment theme. He acknowledged that, *"it is not so well understood that institutional structures and procedures have an equally profound influence on teaching and learning, and the way learning is assessed"* (Haywood 2000:9).

Candy et al. (1994) reported for the Australian government on *Developing Lifelong Learners through Undergraduate Education*, (p149). Their chapter 7 concerned *Teaching and Assessing to Promote Lifelong Learning*. In it, they reported that one of their contributors made the point that:

> *"Since students come to University to gain qualifications, what is in the exam is of ultimate importance. The smart students skip the mass lectures if they are ineffective, accumulate past exam papers, and simply swat up on the most likely questions. Such knowledge is retained long enough to do well in the exam. The real learning at University for such (often highly successful) individuals is how to dissect the system so that one does not waste one's time on irrelevancies".*
> (Candy et al. 1994:149)

It has also been asserted that student assessment was *"a serious and often tragic enterprise"* (Ramsden 1992:181; 2003:13). The author also speculated that it was an important motivator for student learning. Brown (1999a) went even further and warned of assessment difficulties. She asserted that for many students, assessment, presumably summative rather than formative, is a *"nightmare"* (Brown 1999a:3). This is further evidence of the need to address the effects of GSA reliably and sensitively. Above all, there is also the need to treat them consistently.

Entwistle (1996) was another author who summarised the concept of assessment as motivator. He

asserted that *"The single, strongest influence on learning is surely the assessment procedures"* (Entwistle 1996:111). It seems almost to be a plea. He did not specify the age, education stage or socio-economic status of his learners. The book was however concerned with managing independent learning so it was likely to be from a formal education standpoint. In addition, the students were, presumably, mature enough to benefit from independent learning, i.e. they were HE students. This seems to be an important area for future educational research. If the Entwistle assertion is true, then research into improving the situation also needs appropriate resources.

Other researchers who have written on the importance of summative assessment as a student motivator include Snyder (1971), Gammie and Matson (2007) and Brennan (2008).

4.6 Reliability and validity of HE assessment

This section is a review of the literature on reliability and validity as it applies to HE assessment. It is divided into two sections, 4.6.1 is Reliability and 4.6.2 is Validity.

4.6.1 Reliability

The US Government Department of Education *National Postsecondary Education Cooperative* (NPEC) *Sourcebook on Assessment* definition of reliability was that it was an estimate of '*test takers'* performance. It is about their consistency, *"internally, across time, test forms, and raters"* (Erwin 2000:6).

Haywood (1989) asserted that reliability was a problem with summative assessment marking. He noted that the doubt had been *'understood'* since the early twentieth century (1908) in the US and that it had become known in 1935 in Britain. In addition, *"marks of examiners are highly variable"* (Heywood 1989: 51). They were also, he noted, unreliable. (The British reference was probably to the Hartog and Rhodes (1935) study on essay assessment, see section 2.6.)

The Elander (2004) definition was similar to that of the NPEC above. He noted that assessment reliability *"refers to the stability or consistency of marking"* (Elander 2004:36). Most marking research focused on it. It can he noted, be estimated from these marks when more than one is available. He also noted that this happens when work is double marked. This is usually specifically to check the first markers reliability. On the other hand, he also affirmed that validity is a much more difficult issue than reliability.

The reliability of a measure is its internal and external consistency. Internal reliability is expressed in a variety of ways. One of these is the split-half measure, i.e. are all the test items consistent as a sample of the domain. This can be assessed using a widely known statistic, Cronbach's alpha. The second part is the rater consistency. Would the same marker reach the same conclusions about the value of the responses if they carried out their marking again at a different time? External consistency is also in two parts. The first is the generalizability of the measure, i.e. is it applicable in a larger population? The second part of it is inter-rater reliability. Would different markers assign the same marks to the work? (Bryman 2004)

As one of a very limited number of authors commenting on degree classification reliability, Elander (2004) has noted that reliability *"is much higher for degree class than for modules or units of assessment"* (Elander 2004:36). He asserted that this was due to averaging the degree class from across many modules, or *'units of assessment'*. He did not comment on the validity of the degree class.

Another group expressed concerns on the reliability of the UK degree system. Their report noted *"A growing number of studies support Yorke's conclusion that* "the honours degree classification is less reliable than many believe"" (Burgess 2007:25) (Yorke 2007). Unfortunately, neither Burgess nor Yorke offered any additional evidence for their conclusions. The Measuring and Recording Student Achievement Steering Group, of which Burgess was Chair, commissioned a review of the research literature. From it, they concluded that the distribution of degree class and module marks varied with the subject discipline. They also concluded that the assessment method, and the one used to assign the classification, also has an effect on the level of award.

4.6.2 Validity

The validity of the GSA method of marking is at the heart of this present study. The definition of validity in the US Government Department of Education NPEC *Sourcebook on Assessment* (Erwin 2000) was quite succinct. Their view was that it *"involves* "building a case" *that a test is related to the construct it is intended to measure"* (Erwin 2000:8). The report writers also noted that the three types of validity were content, criterion and construct.

Elander (2004) commented on the individual aspects of a student assessment. His was a psychological point of view rather than an educational one and it concerned evidence of the absence of validity. He noted that this was where it could be seen that *"marks were influenced by*

something the assessment was not intended to measure", (Elander 2004:36). He also warned that validity *"is a much more difficult question to deal with than reliability".* (The previous section 4.6.1 is on reliability.) He pointed out that validity, or its absence needs more evidence than just the marks awarded. He noted, *"It is much more difficult, however, to make a direct comparison between marks awarded and what should have been measured, because there is no 'gold standard' for student assessment with which the grades awarded by markers can be compared"* (Elander 2004:36).

It has been noted that "Messick takes the view that adverse impact does not, *itself* render the test invalid" (Gipps 1994:63, emphasis as original). This is, of course, only when construct validity has been shown. Cronbach, on the other hand, argued that adverse social consequences are enough '*of themselves'* to call the validity of a test use into question (Gipps 1994).

Messick's 1989 definition of validity has been widely cited, (e.g. Wiliam and Black 1996:539; Borsboom et al. 2004:1063). It also seems to be part of the language of research methods teaching. He asserted, *"Validity is an integrated evaluative judgement"* (Messick 1993:13). He noted that it concerned the extent to which the evidence and rationale support the assessment conclusion, and the argument from them both seems to be that only the theory and the results are important, not the method.

Another definition of validity comes from Kaplan (1998). Like Messick, his definition was also about consequences. *"The validity of a measurement consists in what it is able to accomplish, or, more accurately, what we are able to do with it"* (Kaplan 1998, 1964:198). Messick, when he cited Kaplan, agreed. He did however also express doubt about such a simple explanation. He asked, *"whether the measures have been so arrived at that they can serve effectively as means to a given end"* (Messick 1980:1012; Kaplan 1998, 1964:198). It appears though, that Kaplan disagreed with Cronbach and Messick's assertions. They claimed that validation applied only to the results from a research instrument, and had nothing to do with the instrument itself. Kaplan insisted that the method was also important. *"the procedure itself plays an important part in specifying the meaning of the term naming the magnitude in question"* (Kaplan 1998, 1964:198).

In the preface of his third edition of his *Essentials of Psychological Testing,* Cronbach (1970) made an important observation about the first edition (1949). He noted that it was *"when everybody who*

was anybody in psychology had recently been a full-time applied psychologist with the military" (Cronbach 1970: xxvii). He added that their role would have been as clinicians, or in some interesting role that they could not discuss.

It has also been noted, *"No test maker can put into his test all desirable qualities. … Tests must be selected for the purpose and situation for which they are used"* (Cronbach 1970:115). Later, he also warned that *"A test that helps in making one decision may have no value at all for another"*, (Cronbach 1970:121). This was also echoed more recently by other authors who added, *" It is certain that no one method of assessment is adequate for testing a course"* (Danili and Reid 2005). Later Cronbach offered some reassurance.

> *"Every test is to some degree impure and very rarely does it mean exactly what its name implies … construct validity is established through a long-continued interplay between observation, reasoning and imagination. … The process of construct validation is the same as that by which scientific theories are developed".*

(Cronbach 1970:142)

He asserted that there were three parts to construct validation. One, *"Suggest what constructs might account for the tests performance. This is an act of imagination based on observation or logical study of the test."* Two *"Derive testable hypotheses from the theory surrounding the construct. This is a purely logical operation."* Three, *"Carry out an empirical study to test one hypothesis after another"* (Cronbach 1970:143). He also commented that:

> *"Validity is high if a test gives the information that the decision maker needs. No matter how satisfactory it is in other respects, a test that measures the wrong thing or that is wrongly interpreted is worthless".*

(Cronbach 1970:121)

Although a degree is not as clearly defined as a psychological construct on which their work was based, several years earlier Cronbach and Meehl (1955) had noted, *"Construct validity is a measure of some attribute or quality which is not "operationally defined"* (Cronbach and Meehl 1955:282). They added that it is *"not to be identified solely by particular investigative procedures but by the orientation of the investigator".* They also noted, *"Construct validity must be investigated wherever no criterion or universe of context is accepted as entirely adequate to define the quality being measured"* (Cronbach and Meehl 1955:282). They also pointed out that in their view, rejecting the null hypothesis does not mean that the construct validity is sound. Researchers must give the fullest possible account of the extent to which they presume the construct validity to be true. The problem for the researcher is not to presume that the test is a perfectly valid method of

measuring the construct. They noted that adequate criteria do not exist for most tests. In addition, they noted that many tests like these remain *'unvalidated'* by their users. Either that or the authors rationalize their method as though it was a validation of it. By this, Cronbach and Meehl remind us that rationalization is not construct validation. They noted that, *"Construct validation was introduced in order to specify types of research required in developing tests for which the conventional views on validation are inappropriate"* (Cronbach and Meehl 1955:299).

Validation includes construct, content, and predictive and concurrent validity. In a 1980 address to the Educational and Psychological Evaluation and Measurement divisions of the American Psychological Association, Messick contended that construct validity "*is* the basic meaning of validity", (Messick 1980:1015, original emphasis). Thus, he reduced at a stroke the types of study validity to one. He cited Guion's 1976 conclusion that all validity is, in its basic form, construct validity. He also asserted that it integrated both criterion and content validity for hypotheses testing. His complex text concluded that construct validity, like so much else in research, is context specific. A further observation at that time was that *'"*criterion-related validity" *is usually considered to comprise two types, concurrent validity and predictive validity"* (Messick 1980:1016).

As long ago as 1933, a warning was issued that *"There are in use today at least one thousand different educational and mental tests"* (Ruch 1933:39). Ruch estimated that less than ten percent had any *'convincing critical and statistical data'* included with them. His warning meant that readers would not be able to judge the validity, reliability or *'norms'* of the measure for themselves. His belief was that both the publisher and the author of the test had an ethical obligation to include such data in the publication. When Ebel (1961) cited this, he added, *"One might reasonably expect that the situation would have improved in the intervening years, but this seems not to have happened"* (Ebel 1961:641). Although it does seem to have at least extended into a more general debate since then.

The construct validity of this study is sound because assessment, both formative and summative, is part of the student cosmos. The data used in this present study had previously been collected by other parties for other purposes. This meant that this present study could have had no effect on the data subjects. The data used were in the form of module marks out of 100. The module mark was from summative assessment items from module components and component elements.

4.7 Portfolios

Koretz et al. have reported,

> *"The rater reliability of portfolio scores in both mathematics and writing was very low. … The percentage of cases in which raters agreed on a score was generally not much higher than expected by chance"*

(Koretz et al. 1993:xix).

In 2001, that is between the Dearing (1997) and the final Burgess (2007) reports, it was noted that perhaps student portfolios have the answer to some of the shortcomings, including the validity and reliability, of more traditional summative assessment techniques. The reasons were summarised in the following quotation.

> *"An essential feature of a portfolio is surely that it comprises a collection of items rather than a single piece of work. In this respect it can be distinguished from a dissertation or project report (although such items may be included in a portfolio)".*

(Baume 2001:7)

When discussing the features of a good portfolio scheme, Baume observed that the issue of Fairness was *'a profoundly subjective concept'* that raised *'strong passions'*. He also noted that validity and reliability contributed to the perception of it. He observed that fairness could not be accommodated into portfolio assessment practice but that a partial and reasonable proxy for it would be to specify a maximum portfolio size (Baume 2001:12). He explained that assessment by portfolio fell short of being the ultimate assessment method:

> *"In suggesting in this section some of the features of a good portfolio assessment scheme, I may have suggested that I see portfolios as the perfect assessment method. I do not hold this view. Assessment of complex skills and knowledge will remain a complex task".*

(Baume 2001:14)

For him, a portfolio was a structured collection. It was comprised of labelled evidence of the student's critical reflection. The intention was to produce it as a part of the student learning process to show evidence of that learning.

4.8 Oppenheim et al. (1967) 'Assumptions Underlying the Use of University Examinations'

In their '*Assumptions',* Oppenheim et al. (1967) observed that *'examination'* had two uses. The first assumption was about past or present attainment. The second was predictive. They made twenty-one assumptions about what they were measuring. One of them was that they expected students to take individual responsibility for their own work.

They presumed that both the purpose and function of examinations were *"related to certain objectives"* (Oppenheim et al. 1967:341). These were educational, idealist and pragmatic. Objectives that they deemed to be *idealist* were those that "*involve the pursuit of knowledge, truth, or beauty"* (Oppenheim et al. 1967:341) for their own sake. Objectives that they deemed pragmatist were those that they thought of as instrumental. This apparently involved the pursuit of truth as a means rather than an end in itself. Their assumptions included those of the consequences of *examinations.* That is, the effect of assessments, on candidates and their parents, and on university staff. Their assumptions on *examination* predictive uses were wide ranging. They included the consequences of the results for several groups of critical stakeholders. These included employers, government, schools and departments, postgraduate schools, professional bodies and university administration. Contribution to knowledge was also in their list of consequences.

They claimed that there were twenty-one commonly held assumptions about assessment. These related to the content and methods of university first-degree *examinations*. This thesis will not review or discuss these item by item. It will review only those that are relevant to this study. These were assumptions 1, 2, 5 and 11. Their Assumption 1, for example, was *"that university examinations can include some so-called imponderables … in their assessment"* (Oppenheim et al. 1967:343). These were qualities of mind, independent critical thinking, and breadth. They emphasized their belief in a dual purpose of assessments. The first was *"to assess the students' knowledge".* The second was to assess their *"quality of mind"* (Oppenheim et al. 1967:343). They also discussed other meanings for these terms. Assumption 2 was that *"'quality' of academic performance is rateable on a single continuum, from first class honours to failure"* (Oppenheim et al. 1967:344). Both of these assumptions relate to construct validity as well as to Newton's point, mentioned earlier that *"The fact that a system which is fit for one purpose will not necessarily be fit for all purposes is a fundamental consideration when evaluating the legitimacy of proposals"* (Newton 2007:149).

Assumption 5 was,

> *"that each examinee should have individual responsibility for his own performance; we do not accept collaboration or team work, no matter how common this may be in real life performance".*
> *(Oppenheim et al. 1967:344)*

They also observed, as part of this assumption, that,

> *"The notion of individual responsibility is typical of our culture at the present time: the idea that we might … perform extremely well together as a group is unacceptable".*
> *(Oppenheim et al. 1967:344)*

Assumption 11 was on trying to prevent bias in assessment by using external examiners. The authors conceded, *"some bias is almost inevitable"* (Oppenheim et al. 1967:345). They acknowledged the human frailty of both university teachers and external examiners. They also restated the external examiners task. It was to judge and attest to the university's standards. One of their concerns was whether the external examiners do help in maintaining those standards. In order to begin to study this, they proposed two questions. The first was the frequency of teacher examiner disagreement. The second was how they are resolved. Elton (2004) also had serious concerns about this. He noted, *"there is extraordinarily little research evidence on the effectiveness of any of the work of external examiners".* He also observed, *"that what there is does not inspire confidence in the system"* (Elton 2004:58).

The Oppenheim et al. assumptions are not an exhaustive list. As the authors acknowledged, and has been pointed out by another author, *"they say nothing explicitly concerning reliability and validity"* (Elton 2004:44). In his report, Elton was reflecting on the formal education system of the 1960's and 1970's. He noted that these assumptions had been re-examined since then. He also noted that many assumptions made then about assessment, also applied to much of contemporary 2004 practice. This had exposed their shortcomings. Elton observed that rather than relating to assessments per se, the failing related to the *"distribution, recording and use of examination results"* (Elton 2004:43). In other words, like Newton (see section4.3), he was warning about the misuse of the results of assessment.

Highlighting the length of time that these assessment issues have been understood in connection with examinations, he also quoted Cox's opening sentence from the same journal issue. *"It is now thirty years since serious doubts were raised about examinations, yet … since then, very little has changed"* (Cox 1967b:352; Elton 2004:33). This would seem to be a clear reference to the work of Hartog and Rhodes in the mid nineteen-thirties (Hartog and Rhodes 1935; Anderson 1953; Hartley 1998; Wilson 1998). On a similar theme, Hill (1972) noted tellingly *"The use of numerical marks in the assessment of examination scripts gives to the assessment a spurious air of certainty"* (Hill 1972:221).

Elton discussed the *Assumptions* in some depth. He added that

> *"To the best of my knowledge, few if any of these assumptions have been seriously addressed by practitioners. Thus, 26 years later Atkins et al. (1993) list six serious and persistent flaws in assessment practice, which include but go beyond many of the concerns of Oppenheim et al."*
> (Elton 2004:43)

He noted that these flaws included and exceeded those assumptions of Oppenheim et al. The Atkins et al. flaws are listed in Appendix 8. While they may be somewhat critical of academics in general, their underlying points would appear to be difficult to refute. Elton commented, *"I submit that the time has come to take a serious look at the issues raised"* (Elton 2004:44). He continued by noting that the drivers of assessment also included the Quality Assurance Agency (QAA), rather than academics alone.

At the time of the Oppenheim et al. *Assumptions* (1967), summative assessment of undergraduates would have been synonymous with individual written examination. No published reports found on GSA from that period. The implication is clear. In 1967, the concept of GSA was unknown.

In addition, for Oppenheim et al., there was an implicit assumption that, for example, there will be few, if any, self-financing students. Neither will there be any otherwise independent or mature students. There was no explicit *return on investment* on their examination purposes clause 1(a). On the other hand, there was in their Assumption1 (c). The former is about the students themselves. The latter is about their parents. Today their *imponderables,* i.e. quality of mind, independent critical thinking, and breadth, in their Assumption 1, might be included in the term graduateness. Assumption 2, predictive examination purposes, seems broadly, to be widely held. Otherwise, why is the undergraduate degree classification system still in use today? This should improve post Dearing and Burgess, as higher education institutions include graduate transcripts (HEARs) with the degree classification parchment assuming they materialise.

If examination was a 1960's term for assessment, then Assumption 5, that each examinee should have individual responsibility for his own work, is especially important to this present study and to formal education more generally. It is however explicitly incompatible with GSA. Since then, andragogical theory seems to have moved away from the second part of Assumption 5. Assessment in HE now embraces group work.

4.9 The joint AERA, APA and NCME 'Standards' fairness

The observations on fairness in the second edition of the *Standards* (Baker et al. 1999) reflect much of the criticism of GSA.

The Committee to Develop Standards for the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME) jointly prepared the *Standards for Educational and Psychological Testing*, (Baker et al. 1999). In their opening paragraph they noted that *"The improper use of tests, however, can cause considerable harm to test takers and other parties affected by test-based decisions"* (Baker et al. 1999:1). Also, see Cronbach (1970:152) in section 4.10 and further discussion on the issue of harm to *'test takers'* and others in section 8.2.3.

The opening lines of the introduction to the earlier 1985 edition of The *Standards* seemed, to this author, more detailed in its observation. It was that

> *"Educational and psychological testing represents one of the most important contributions of behavioral science to our society. It has provided fundamental and significant improvements over previous practices in industry, government, and education. It has provided a tool for broader and more equitable access to education and employment. Although not all tests are well-developed, nor are all testing practices wise and beneficial, available evidence supports the judgment of the Committee … that the proper use of well-constructed and validated tests provides a better basis for making some important decisions about individuals and programs than would otherwise be available."*

(Novick 1985:1)

This same edition of the *Standards* also stated that as a general principal of their use

> *"The test user, in selecting or interpreting a test, should know the purpose of the testing and the probable consequences. The user should know the procedures necessary to facilitate effectiveness and to reduce bias in test use."*

(Novick 1985:41)

The later edition (Baker et al. 1999) seems to have incorporated this principle into two of its standards. These are in their book sections on the responsibilities of test users (in standard 11) and in their educational testing and assessment section (13). These standards were applicable to all evaluation. They included test instruments produced by third parties under *'contractual arrangement'*. The 1999 edition specifically included an additional section entitled *Fairness*. The *Standards for Educational and Psychological Testing* detailed four principal ways in which the term is used, which are reviewed following this introduction. (The *Standards* is also cited in section 2.7.3.)

4.9.1 Fairness as lack of bias

The *Standards* explained that it uses the word *Fairness* as a technical term. The authors asserted that the problem arises when *"deficiencies in a test itself or the manner in which it is used result in different meanings for scores earned by members of different identifiable subgroups" (Baker et al. 1999:74).* This, they reminded us, is differential item functioning (DIF) (also see section 2.7.3). At the same time, they noted that *"There is a general consensus that consideration of bias is critical to sound testing practice" (Baker et al. 1999:74).*

4.9.2 Fairness as equitable treatment in the testing process

The *Standards* authors also asserted that there is a consensus that *"just treatment throughout the testing process is a necessary condition for test fairness"* (Baker et al. 1999:74). They continue the theme by adding that its requirements include *"not only of a test itself, but also the context and purpose of testing and the manner in which the test scores are used".* The authors argue that there is no intrinsic fairness or unfairness in the design of a good test (i.e. assessment). Rather, it is the use of the test, and the circumstances surrounding it, which impose either of these qualities upon it.

4.9.3 Fairness as equality in outcome of testing

They report that there is no support for the concept that for fairness to be the same across different assessment groups, the marks distributions should be similar. They assert that although marks distribution differences across these groups should be investigated, they *"do not in themselves indicate that a testing application is biased or unfair"* (Baker et al. 1999:75).

4.9.4 Fairness as opportunity to learn

The principle of a fair opportunity to learn means equal access opportunity for all, irrespective of their ability to pay, or other access issues. The *Standards* reported, *"At least three important difficulties arise with this conception of fairness"* (Baker et al. 1999:76). The first was the practical difficulty in developing a suitable definition of the term opportunity to learn. Second, they observed that even documenting the curriculum topics was difficult, and doing the same for each student *"may be impossible".* Third, they commented on the *"well-founded desire to assure that credentials attest to certain proficiencies or capabilities"* (Baker et al. 1999:76).

An educational award, e.g. a degree or diploma, is in recognition of proficiency. They should not be awarded, for example, simply because the candidate claimed not to have had time to learn the material. The claim would need to be judged on the details. To pass such a candidate could mean *'certificating'* someone as proficient who may not be.

4.10 The importance of summative assessment in higher education

As mentioned at the start of this chapter, several authors and organizations have published work on the importance of assessment in higher education. This importance is enshrined in the oft-cited observation of Boud (1995b) in the introduction to Chapter 2. Also, see previous section, the focus of which is the *Standards for Educational and Psychological Testing* interpretation of fairness, (Baker et al. 1999:73).

It has also been asserted that *"assessment is at the heart of the undergraduate experience"* (Brown and Knight 1994:12). This was also cited by Brown et al. (1997a). They observed that it defines that which students regard as important in relation to their time at university. These authors had no doubt about the importance of assessment in higher education. It was explained that *"It is a legitimate concern of those who learn, those who teach and those who are responsible for the development and accreditation of courses"* (Brown et al. 1997a:7). To them it was *'the cash nexus'* of learning. Students take their cues about assessment from the topics to be assessed. What teachers said was important for students, was not always what students thought was important to themselves. It has been noted that *"Put rather starkly: If you want to change student learning then change the method of assessment"* (Brown et al. 1997a:7). The authors then briefly outlined the work of several groups of authors whose late 80's and early 90's work informed their assertion. For example, they stated the belief that multiple choice test assessment was associated with reproductive styles of learning, i.e. surface learning, and that, in their view, coursework projects promoted independence and deeper strategies of understanding.

It has also been asserted that in its most fundamental form, the assessment process consisted of taking a sample of the work that students do, and then judging it (Brown et al. 1997a). The assessors then made inferences about it. The process ended with an estimate of worth in the form of grades, marks, degree classification or recommendation. The point has been made that *"All forms of assessment provide estimates of the person's current status"* (Brown et al. 1997a:9). This emphasised that the sample may not be entirely generalizable to the student. Brown et al. was another group of authors who asserted that assessment results have two prime uses. For them, one was judgmental, or summative. The other was developmental, or formative. Judgmental assessment was concerned with licensing the student to proceed to the next stage, either to an appropriate level in their studies, or to graduation with an appropriate classification. It was

concerned with accountability strategies and its hallmarks are *consistency, uniformity* and *fairness*. In a statement of the purposes of assessment, Brown (2001) noted that they were

> *"to give a licence to proceed to the next stage or to graduation; to classify the performance of students in rank order; to improve their learning. Differences in purpose should influence the choice of assessment task".*

(Brown 2001:3)

He also confirmed that the assessment method should always be specific to the assessment context. In his collaboration with Much in 2001, they expanded this purpose into three categories: learning, certification and quality assurance.

This Paulsen (1908) passage, also cited by Elton (2004), highlights the influence that assessment exerts on student behaviour, and the dilemma this presents to them.

> *"The prospective examination necessarily turns the student's attention from the subject itself to the examination that must be passed. This leads to appreciable disturbances of scholarly study ... and still further injury is done by the fact that the examinations compel him to busy himself with things that are of no real value to him and thus prevent him from following his inclinations for other things".*

(Paulsen 1908:341; Elton 2004:55)

An anonymous submission to the Candy et al. (1994) committee led them to note in their report, *"Since students come to university to gain a qualification, what is in the exam is of ultimate importance"* (Candy et al. 1994:149). The contributor commented, succinctly, *"The tyranny of the exam is captured in this submission".* Entwistle emphatically endorsed this belief when he asserted, *"The single, strongest influence on learning is surely the assessment procedure"* (Entwistle 1996:111). He also asserted *"Higher education does not involve certification of competence"* (Entwistle 1996:111). Both comments are difficult to understand. They are particularly difficult given his *strongest influence* assertion. This belief however, was further endorsed by Boud et al. who argued, *"assessment is the single most powerful influence on learning in formal courses"* (Boud et al. 2001:67). Other authors have made a similar observation, e.g. Thompson and McGregor (2005). There cannot be much doubt that it is a strong influence on learning. It is difficult however to understand how these researchers could all be so sure that it was *the* single strongest influence.

Another proponent of the importance of assessment is Stobart (2008). He noted, *"the quality of the summative assessment on any course will strongly influence the learning approach"* (Stobart 2008:85). Other authors have later reminded us, *"the best student found by one method is not necessarily the best student found by another method"* (Danili and Reid 2005:204). These authors

asserted that thus are questions raised concerning the validity of the assessment formats, especially what they are measuring. Other authors had earlier raised the point that assessment needs to be context specific. They explained that *"the design and the methods of assessment ... need to take into account the main purposes of education"* (Miller et al. 1998:20).

Assessment might not just affect student behaviour. It might also affect educational policy. Yero quoted the business guru Tom Peters. *"What gets measured, gets done"* (Yero 2002:3). She also noted, "Regardless of high-sounding rhetoric about the development of the total child, *it is the content of assessments that largely drives education*", (original emphasis).

For Cronbach, assessment, or *'observation'*, was too important to be left to only one occasion. In his warning, he included that, where possible, modules should also be assessed by more than one method. He noted, *"No single observation is entirely representative of the person"* (Cronbach 1970:151). How many times should we assess students? Is twice enough, or are three times too many? What are the implications of his advocated change of policy in terms of student's time and university resources? There was no discussion. This would seem to be a future study opportunity. He went on to state that *"An erroneous favourable decision may be irreversible and may harm the person or the community"* (Cronbach 1970:152). (Also see Baker et al. (1999:1) in section 4.9.) As he noted, *"Even when reversible, an erroneous unfavourable decision is unjust, disrupts the person's moral, and perhaps retards his development. Research too, requires dependable measurement"* (Cronbach 1970:152). It has important implications for all stakeholders of degree classifications in general, not only those of GSA.

4.10.1 Some examples of GSA practice

It has previously been noted that groupwork is a popular teaching, learning, and assessment method. Brown and McIlroy, for example, noted that *"Group learning activities (GLAs) are integral components of graduate and undergraduate programmes across disciplines"* (Brown and McIlroy 2011:687). A consequence of this is that GSA must be used equally frequently see section 4.5.6.

The usual group comprises four to six members. They have been found used in the full gamut of disciplines from Archaeology to Zoology. For example, Durham University's archaeology practical modules used it (Durham University 2006). At least one zoology programme used it (University of Western Cape 2004). Their webpage used the term group-work. It was even mentioned, as group learning, on the University of Central Lancashire Department of Physics, Astronomy and

Mathematics, Extraterrestrial Life module website (University of Central Lancashire 2004). Other, more conventional uses, as an example in disciplines beginning with the letter M, are in Table 10.

Table 10: Some examples of GSA other module disciplines

Marketing (Gatfield 1999; Kates 2002)
Management (DeVita 2001)
Maths (MacBean et al. 2004) and (Townend 1997)
Medicine (Helsinki 2004)
Microbiology (UCL 2004)
Modern Languages (Keele 2005)
Music (Ulster 2004) and (York 2002)

It has been observed that, *"universities around the globe have recognised the need for training of undergraduates in teamwork skills"* (Murray 2003). In addition, as mentioned earlier (e.g. see section 2.7.2), the literature abounds with methods of forming student project groups fairly and transparently. There are also studies of methods of administering and delivering them.

A student's assessment mark is important because it is measurable and understandable. It is also, a comparable and visible recognition of the time, effort and ability they have given to the project (Wellington 2004). It has also been noted that: *"Authorities on assessment draw attention to the need to match assessment with teaching and learning objectives … and the link between assessment and student learning"* (Simonite 2003a:468).

4.11 HE assessment literature chapter summary

This chapter has highlighted the need for further research to address the questions in table 1. Previous work on summative assessment in higher education was reviewed. Section 4.2 reviewed the purposes of higher education. Section 4.3 reviewed the roles and purposes of assessment in higher education. Section 4.4 concerned the role and definition of summative assessment. Section 4.5 was about GSA use. Section 4.6 concerned the reliability and validity of HE assessment. Section 4.7 concerned Portfolios. Section 4.8 reviewed some of the Oppenheim et al. assumptions, underlying the use of university examinations that were relevant to this study. Section 4.9 reviewed The *Standards* views on Fairness. Section 4.10 illustrated the importance of assessment.

In common with the general background literature review in the previous chapter, published work on HE summative assessment also appears to have missed the opportunity to report the impact of groupwork on the overall student mark, and therefore its impact on their award classification. In terms of GSA, it appears to have concentrated more on methods of deriving individual marks for

GSA projects. This study will begin to close the gap in the literature on the impact that GSA may have on student overall marks.

The next chapter (5) will highlight the study methodological and design considerations.

Chapter 5. Study design considerations: Method and methodology

The last two chapters reviewed the literature that provided the background to this present study and showed a need to address the study questions presented in table 1. In this chapter, the study methodology, method and design are considered.

The concept and practice of summative assessment is familiar to the student population. It will have been applied to them throughout their schooling. Summative assessment, even as a group, is therefore unlikely to be entirely novel to undergraduates. It is just another variant of summative assessment. From this, most students assessed by GSA should be able to grasp the nuances of its collaborative and collective nature compared to that of IISA.

This present study was not in any sense phenomenological, i.e. it was not from the point of view of students' *experiences* of GSA, although this would be an excellent approach for a future study. In addition, it had no influence over the data or the outcomes of the data subjects studies. The students will be unaware of this present study. It analysed secondary data from HEI official records. This was a retrospective, post-positivist survey. It was not an intervention study. Positivists take the epistemological stance that it is acceptable to apply the empirical methods of the physical sciences to topics belonging to the social sciences, e.g. to educational issues. They assume that it is possible to conduct studies on social science topics that are both objective and value-free. Post-positivists, on the other hand, accept the imperfections in the world. They address the issue of researchers bringing their own experiences to their research, and try to reduce their impact (Bryman 2004).

As mentioned elsewhere in this thesis, (e.g. section 2.2), there is no single, named, unifying theory of GSA upon which to base this present study design. This lack is most noticeable in that there are, instead, many methods and variations of methods of deriving individual marks from group marks (see section 4.5.5 and Appendix 5). Academic practitioners of GSA seem to take a pragmatic approach when devising marking schemes. In addition, it seems possible that including a GSA element in a module-marking scheme will not always be for educational reasons. It may be included for compliance reasons or sometimes it seems, even for cosmetic or aesthetic reasons.

Each individual graduate receives an individual academic degree classification. The award is for

the successful end to a degree programme from a degree awarding institution. It derives from their aggregate module marks. These are a measure, however imprecise, of the individual holder's attainment in the assessment, i.e. their graduateness (see section 2.7.1). This present study has relied on existing student marks records. The data analysis was also dependent on sufficient detail being accessible from the existing student marks records. (Also, see section 2.3 and 2.7.3).

An ideal study method would have been to explore a random sample of student marks data from the whole student population. This would have meant taking the random sample from a sampling frame (Bryman 2004; Gray 2004). One of the first questions for consideration when establishing a sample frame is what is the study population. For this present study for example, was it only to be first-degree students at one university in the northeast of England, or every student in the world, or some compromise between the two? A study design based on the second option sample frame would be unrealistic, even for a team of fully funded researchers, and even more so for one lone, self-funded researcher.

The scale of the sampling problem exists alongside two main study design problems for a study such as this. While the first of these is defining the population, the second problem is one of data access. Even when the latter is resolved, defining the population could be daunting. For example, new higher education institutions seem to appear with growing frequency in the UK. Some may even disappear in the wake of the university funding reorganisation. This could cause rolling difficulty in setting up and defining the study population. It could cause problems in establishing a suitable cut off point for adding new educational institutions to the sampling frame. Longitudinal data could even be from a different group of institutions each year. In the wake of all of this, it was decided not to base study data on a random sample from the student module marks data population. They were collected from a convenience or happenstance sample. They were also, to some extent at least, a snowball sample. They were what were available to this researcher, at the time of the study.

The pilot study for this research was reported first in Almond (2006) and subsequently in Almond (2009). One of the options for this main study was for it to have extended the pilot study. The intention had been only to use data from two main sources in a university in the northeast of England, and then the opportunity for including some additional data arose. The new study was to have included larger, longitudinal data sets of student marks, and some limited biographical data.

The criterion for including modules in the present study was that the student had studied both IISA and GSA modules and that their marks were available at least to module level, if not necessarily to assessment item level. In addition, a significant part (>5%) of at least one of their module assessments must have been by a GSA method. In this present study, GSA and the IISA marks were on the same scale. They were usually a score between 0 and 100.

The software applications used in the preparation of this thesis, and for the data analysis and the graphics, were Microsoft Word and Excel 2007 and 2010, SPSS version 15.0, Minitab version 14, MIX 1.7 for the meta-analysis, and MLwiN version 2.15 for the multi-level analysis. In addition, some of the figures were drawn using Microsoft Paint software. The regression analysis and multilevel modelling are in sections 7.4 and 7.5. The meta-analysis is in section 7.6 (The limitations of the study are in section 8.2.14.)

This chapter has seven sections. The first, 5.1 is on the subject of the study sample data. Section 5.2 introduces hierarchical data structures, i.e. multilevel modelling, and section 5.3 considers simulation. Section 5.4 illustrates the three types of variables that were relevant to the study. They were independent, intervening (e.g. the students motivation and ability,) and dependent variables. Section 5.5 considers the GSA regulations that applied at one university in the northeast of England at the time of the study. Section 5.6 outlines the data analysis strategy. The final section summarises the methodology, method and design considerations.

5.1 The data sample sources

The data are described in section 6.3 and analysed in Chapter 7.

5.1.1 Source of data samples A1-5

After the present study had started, during 2008, two additional happenstance data sets became available. The first was through a serendipitous meeting with an overseas academic who was also a university teacher and convenor of a suite of GSA modules. They were on a research sabbatical visit to this researcher's workplace. The data were from students studying at an Australian university. Data sets A1 to A5 were also the only ones from postgraduate students' module marks, rather than from undergraduates.

5.1.2 Source of data sample B6

The second happenstance data source was from Figure 7 of Knight (2004), duplicated here as Figure 6.

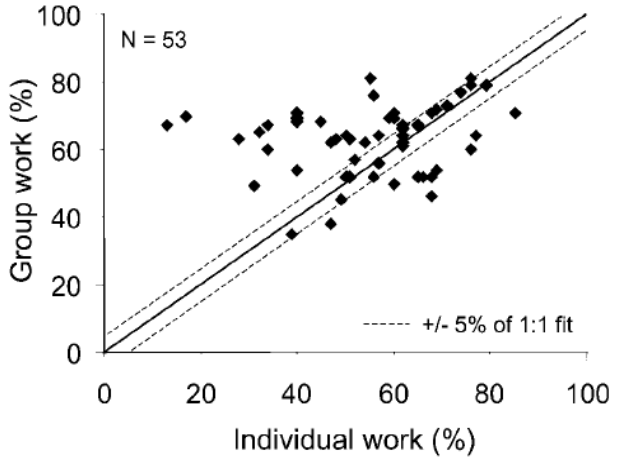


Figure 7. The relationship between students' marks attained in the individual and group assessed exercises. *Note:* The solid line shows a 1:1 fit; the dotted lines represent a +/- 5% total variation from this fit

Figure 6: Figure 7 in Knight (2004)

Data set B was produced by synthesizing it from the chart in his article. This was done with the authors approval (Knight 2008b). The synthesized data is in the table in Appendix 9.

5.1.3 Source of data samples C7-10

Data source C was a programme from the Science Faculty of a university in the northeast of England. The data collected covered study years 2004-5, 2005-6, 2006-7 and 2007-8 and the data sets are labelled C7 to C10 in this present study. The marking hierarchy, and an illustration of how the terms *module*, *component* and *element* are interpreted in the C data set in this study is illustrated in Figure 7.

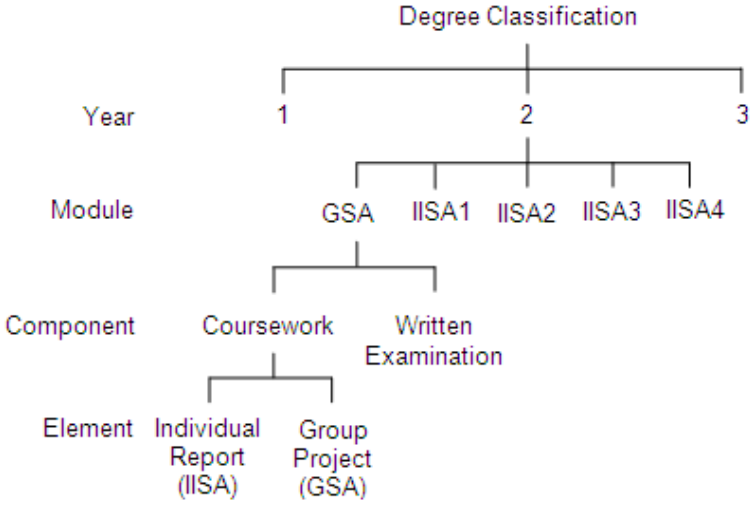


Figure 7: Data source C summative assessment marking format

Marks data was collected at the element level. In addition, all the requested student detail was included in the data supplied. This allowed a greater depth of analysis and a more detailed analysis of the quantitative impact of GSA marking within one programme at this university.

In addition, it had quickly become apparent during the present study that as well as assessing student attainment in the module topic, an important additional aim of the GSA module in the C data degree programme was to provide the students with practical experience of group-dynamics.

As mentioned elsewhere, this was one of only two modules found that simulated the real world scenario of work group non-self-selection. In this one, an administrator applied an algorithm that was designed to result in mixed academic ability groups based on their prior attainment. This method is detailed in Appendix 10.

5.1.4 Source of data samples D11-18

The search for a suitable source of data for this present study began with a search of the module description web pages of the HEI On-line Students' Handbook from a university in the northeast of England, for the terms *group work, groupwork, team work* and *teamwork.* This resulted in data source D.

Several further searches were made to check for omissions from previous searches and for updates to the module list. Some of these additional checks specifically included perusing the module aims and key skills sections. This was done where the module name might suggest its suitability for GSA, for example, laboratory or field based practical modules or those requiring a finished product to be marked rather than just a written report.

Where the module description of the summative assessment method only listed examination and/or essay, then the assumption was that the module items were all assessed as individual work. The on-line searches identified forty-five potential GSA modules that might be suitable for this study. Thirty-one modules were included following semi-structured interviews with the convenors of the forty-five modules (see Convenor interview schedule, Appendix 11).

The second non-self-selected group was included in this data set (the first was noted previously in section 5.1.3). The module convenor attempted to ensure that the members of each group had an appropriate mix of practical project skills, as well as academic ability, between them. This was based on the cohort self-reporting their skills on a pro-forma. All other module marks data in this study were from self-selected groups. The convenor allocated students to produce an appropriate mix of the skills necessary for the group project. The administrator role was similar to senior management allocating the employee project team members. Although the module had only been

running for a few years, this group allocation method was reported by the convenor to have remained virtually unchanged, and trouble free, since the GSA module began. This module model is now (2012) being piloted in other UK universities.

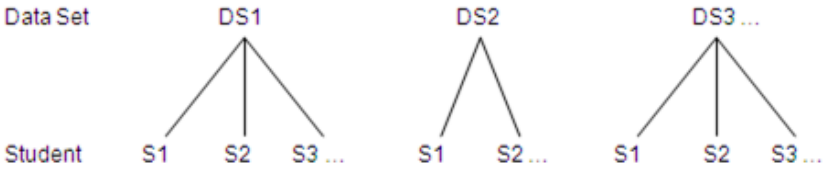
5.2 Hierarchical data structures and Multilevel Modelling

Hierarchically structured (multilevel) data are common in the social sciences, e.g. education and psychology (Dedrick et al. 2009). In the preface to their *Introducing Multilevel Modelling*, Kreft and De Leeuw (1998) note *"Statistical methods are always imperfect tools for achieving understanding of a complex world"* (Kreft and Leeuw 1998). It has also been noted, *"in the real world data are often hierarchical. This just means that some variables are clustered or nested within other variables"* (Field 2009:726).

Multilevel modelling is a data analysis technique that avoids, or at least reduces the risk of, Type-1 errors, (Peugh 2010:85). A Type-1 error is one of falsely rejecting the null hypothesis, (e.g. Clegg 2002; Field 2009). A Type-2 error, on the other hand, is one of not rejecting the null hypothesis when it is false.

We can consider formal education as an example of a hierarchical organization structure. Compulsory state run education in the UK provides a clear illustration. Pupils are taught in classes. In turn, the classes are in schools and the schools are part of the LEA.

Typically, but not always, a multilevel model (MLM) would assign people, in this case pupils or students, to level one (Rasbash et al. 2009). Classes would be at level two for this example. They are in schools. Schools would be at level three of the model. The LEAs administer schools. The LEA would be level four of a hierarchical or multilevel model. In the UK, schools could be further categorised into private and state run, forming a fifth level. Were the study to have an international sampling frame, then the country the pupil was at school in would be level six. Complex MLMs can also accommodate cross-classified data and multiple-membership. These models can contain pupils belonging to more than one class, taught by different teachers, and even taught at more than one school. In these models, people are not always at level one. More generally, units at one level nest within units at the next higher level. The hierarchical structure of this present study data has only two levels. The students are one and the data set is level two. This structure is illustrated in the Unit Diagram, after Rasbash (2008), in Figure 8.



Unit Diagram of the Two-Level Nested Study Data Structure: Students in Data Sets
Rasbash (2008)

Figure 8: Two-level MLM study data structure

The TRAMSS (Teaching Resources and Materials for Social Scientists) website is ESRC funded. In it the authors point out that regression analysis estimates the relationships between response and predictor. They also point out the logic error in applying it to hierarchical data. *"A fundamental assumption of this regression model is that the residuals … are independent. However, data often have a multilevel structure which violates this assumption"* (TRAMSS 1999).

The MLM technique also assumes that the data is a sample from the population. There will often only be a sample available, rather than the whole population. As has been pointed out, *"it is the population of groups from which our sample was drawn which is of interest"* (Rasbash et al. 2009:26). The eighteen data sets for this study are a happenstance, or convenience sample. They are not a random or even pseudo random sample from the population. (Also, see the data sets descriptions in the next chapter). Nevertheless, the analysis was conducted assuming that the sample was representative of the population.

Interest in using MLM for data analysis arose from reading the parallel work done by Simonite. (Simonite 2000, 2001, 2002, 2003a, 2003b; Simonite and Browne 2003c) Her study closely mirrored this one. She used it to analyse the effect of coursework marks on students degree classification compared to the effect of their examination marks. Her dependent assessment category was coursework in general rather than groupwork in particular. It was a study of the effect of one category of assessment marks on students overall marks. In her study, the independent variable was the examination score, i.e. it was the IISA variable rather than the GSA. On the other hand, this present study focused on the link between students coursework GSA module marks on their overall marks.

Not all of the students in Simonite's studies benefited from coursework. She estimated that for 20% of them *"coursework assessment is associated with lower mean marks than examination*

assessment". (Simonite 2001:268). This estimate supports the findings of this study.

5.3 Simulation considerations

This section considers the implications of data simulation. The robustness of the regression analysis finding was explored. The exercise was decided upon after observing the distribution pattern of the dual-line charts described later (also, see Figure 26, section 7.4.3 and Appendix 12). The pattern of the bisecting regression lines and the gentle slope of the GSA regression line compared to the IISA line slope were of special importance. Simulations were set up to compare the extent of any similarity between the present study data scatter plots of correlated study data with those produced from uncorrelated simulated data. It was also to test the alternative hypothesis $H_{2,}$ that uncorrelated data will show the same pattern in a dual-line regression chart as the study data (repeated from Table 1). A purely nominal sample size (100), means (IISA 50, GSA 60), and SDs (IISA 10, GSA5) was chosen.

There is further analysis and discussion on data simulation in section 7.7. Unlike the data in Figure 40 to Figure 42, in section 7.7, for example, the C and D data pairs were not random. Each pair of study data was from the same student.

5.4 Extraneous, mediating or intervening variables

This section considers extraneous, mediating or intervening variables (Creswell 2003; Gray 2004) and the potential they may have to impact on students' overall marks. They include group and student attributes other than the module assessment method and overall mark. In the model, independent variables are the GSA item marks and the IISA item marks. The dependent variable is the module mark for the individual student. Figure 9 shows the relationship between variables.

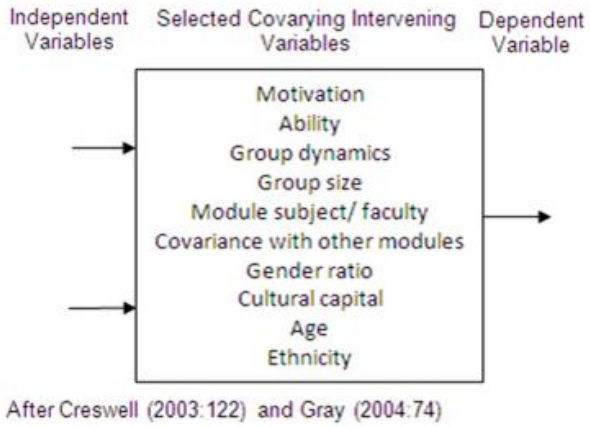


Figure 9: The relationship between independent, intervening and dependent variables

There are other intervening variables, more personal to the student, in addition to those noted by Creswell and Gray in Figure 9. It seems likely that the intervening variables in this present study

also include those in Table 11, are mostly unknowable, and most of them will vary with time.

Table 11: Additional intervening variables

1	General and specific state of physical and mental health
2	Personality type
3	Confidence level
4	Assessment anxiety level and coping strategy
5	Current and future employment (or PG study) ambitions
6	Perceived parental expectations
7	Relationships with their peers, teachers, university etc.
8	Financial circumstances

These additional intervening variables will increase the complexity of the relationship between the IISA and GSA, marks and students' overall marks. In this present study, they cannot be controlled for and they have an unknown impact on the dependent variable.

5.5 HEI groupwork regulations

The lack of uniformity in HEI regulations concerning groupwork assessment was considered, and reflected upon at length, during the design preparation of this study. Table 12 shows a section from a policy note that was part of one university's 2002 groupwork regulations.

Table 12: Part of a university regulation on groupwork assessment

"1. The obvious problem in assessing group work is that of being fair to the individuals within the group not all of whom may have made an equal contribution to the group task. If group work is to be assessed summatively the following examples of good practice should be considered".

(Durham 2002: section 3)

It ruled only that *'If group work is to be assessed summatively …'* In other words, it was not mandatory to assess group work summatively. It seems that the writers did not appreciate the necessity of summative assessment as a requirement for student engagement with the topic, see section 4.5.6. They did however seem to understand the fairness issue with GSA practice. Assessing group work fairly means being fair to the individuals in the group as well as to the group.

In addition, it has also been reported in 1997 that around a third of old universities did not allow group assessment to contribute to students' overall marks in years one or three (Lejk et al. 1997:84). It may have been because some universities' assessment results in the final year have a greater weighting towards the degree classification award. The authors also noted that two HEIs did not allow GSA *"in any part of their chosen course"* (Lejk et al. 1997:84).

5.6 Data analysis strategy

A variety of different analyses is employed to look at the same large and complex data sets. Each method takes different assumptions and a different approach. If all the analyses come to similar or identical conclusions then their validity is enhanced. If not then contradictions can be probed

further.

As indicated elsewhere, correlation between the data categories lies at the heart of this thesis. Unless non-parametric techniques are used correlation analysis assumes that data is normally distributed.

Simple regression ignores the hierarchical nature of educational data. On the other hand, this method is, widely understood and produces results that are readily interpreted.

Multilevel modelling addressed the hierarchical nature of the information but assumes that the data are a random sample from a population. It addresses the shortcomings of simple regression by including appropriate error terms. However, it is a more complex analysis to apply and interpret.

Meta analysis assumes that each data set is from a separate experiment. It combines the data from all of these 'separate experiments' to estimate the robustness. Part of the output is in the form of a forest plot (e.g. Figure 37). It does not need to assume normality

Simulation was added to answer the question: if we assume that the test results were due to certain factors (a theoretical model) can we mimic the empirical data with artificially produced data corresponding to the model?

5.7 Summary of methodological and design considerations chapter

This chapter has highlighted the methodological and design considerations required for the present study. It included an introduction to the study data sample (section 5.1), hierarchical data structures and MLM (5.2), and simulation considerations (5.3). Section 5.4 illustrated known and unknown independent, intervening and dependent variables impacting on students' marks. Section 5.5 illustrated inconsistencies in some university regulations in relation to GSA. Section 5.6 was an outline of the data analysis strategy.

This present study will add to the GSA research thread. It will raise awareness and improve knowledge of the quantitative impact that GSA has on the degree classifications awarded to students assessed, at least in part, by this method. The next chapter (6) concerns the data and the data collection.

Chapter 6. Data description and collection

Methodological and design considerations were discussed in the previous chapter. This chapter describes the data and the main data collection issues.

Students' overall marks, which led to their degree classification, were derived from their year marks. In turn, these year marks were derived from summative assessment marks awarded to the individual student from qualifying modules studied during that academic year. The module marks were derived from coursework and examination assessment component marks. Where the coursework was identified as such in the data received, it was further divided into assessment item components for both individual and group work. An example of module marks hierarchy, for data source C, is shown in Figure 7.

The data were not received in a consistent format from all four sources. The IISA and GSA marks data were received in three different formats:

> The most detailed marks data were from the A and C data sets. The A data were from GSA modules. The C data were from all the students modules for that study year. Coursework component element marks were collected from each.
>
> The B data set source was a single GSA module. IISA and GSA component marks were derived from it.
>
> The least detailed data set was D. It contained student marks for individual modules that had been previously designated for this study as either IISA or GSA. These marks were averaged to produce two variables, the IISA modules mean mark and the GSA modules mean mark for each student.

Assessment marks attempt to quantify student attainment in some aspects of the module topic. It follows from this, that separate IISA modules should measure the individual student attainment in the respective topics. Aims and outcomes are specific to the individual module. In addition to the group product report on the module topic, and/or a project product, i.e. whatever was being assessed, a separate individual report was often included as a summative assessment item in the GSA modules. This provided an opportunity for the student to comment on the group or on individual behaviour during the project.

For most of the student data (i.e. data from source D) used in this present study, no biographical data were available. Some limited details were included in data sets A and C, e.g. age and/or gender. In addition, data from source C included entrance qualifications for some, although not all candidates, and where available, degree classification, (see section 7.2.3). The entrance

qualifications data were inconsistent and incomplete so further analysis was not possible in this present study.

All marks data included in this present study allowed a comparison of students IISA scores with their GSA scores. The data were from students who studied their GSA modules between 2004 and 2008 inclusive.

Module inclusion in the present study depended on there being marks from both GSA and IISA modules for each student. The data were mostly from undergraduate degree modules

Typically, the GSA module data in this present study came from two to five assessment items. There was only one GSA item in each GSA module. This typically contributed between 15% and 50% of the total module summative assessment mark. Exceptionally, one GSA module GSA item had a weighting of 85% of the total module marks (see sections 3.5.4, and 8.1.3).

There was very little consistency in GSA marking criteria across modules, disciplines, schools, faculties or higher education institutions. For example, the way the marks were split between IISA and GSA elements, and the number of assessment elements that were included in the assessment component, (see Figure 7 for example).

Section 6.1 explains why undergraduate data were preferred for this present study. Section 6.2 is on the ethical considerations of the present study. Section 6.3 deals with Data description. The dilution effect that IISA item marks could have on GSA module marks from the D data set is discussed in section 6.3.4.1.

6.1 Why undergraduate data were preferred for this study

Selecting undergraduates and graduates as data subjects, rather than post-graduates, was the logical approach given the origin of the research problem (see following and section 1.1). Why the original intention had been to limit the data, rather than include all taught degrees to undergraduate and graduate data subjects, is a different question to the one posed earlier about why study the topic at all (see section 2.7). The answer is as mundane as it is pragmatic. This researcher felt the emotional impact of GSA most intensely while studying for a Bachelor's degree. During the time that the author studied his first Master's programme, which immediately followed, the impact was

benign by comparison.

In addition, the undergraduate population is greater than that of postgraduates. There were, for example, more undergraduates than there were post-graduate students at HEIs in the UK in 2008/9, in a ratio of around 3:1, (UUK 2010:6) at the time of the present study. A note in one university brochure indicated that there were almost eleven and a half thousand undergraduates compared to just over four thousand postgraduates, (DU_2009Brochure 2009:48; 2010). For this exploration, it was more likely that there were more suitable undergraduate modules than taught postgraduate ones.

As previously mentioned, early in the present study there was a meeting between this researcher and a senior academic from an Australian university. They had an overlapping interest in the effects of GSA marking. This serendipitous event resulted in the inclusion of PG student data into the study. (Data source A).

6.2 Ethical considerations

Ethical permission for this study was sought and received from the Ethics Committee of the School of Education, Durham University. A copy of the text of the permission e-mail is in Appendix 13.

There are two main ethical issues with this present study. The first is that no student should be identifiable from any published account of the study. The second concerns the raw data. It is to be kept secure during the study and destroyed it at its conclusion.

As was mentioned in the chapter introduction, this study is retrospective. It utilises secondary, existing, university student marks data, therefore there could be no Hawthorn effect. Neither could the progress of this study, or its outcome, affect the degree classifications of data subjects.

Additionally, the data sets were anonymised. There were several reasons for this. One was to discourage possible future attempts to identify the students, programmes or modules directly. Others were to comply with research ethics, as a condition set by the Australian (data source A) academic of being granted access to the data, and to comply with data protection law. The source of the B data (Knight 2004) was already in the public domain and students were already anonymised in the article.

The anonymised C data were from the amalgamation of several separate data files from one

school/department. Another data set, where anonymity was a condition of access, was from data source D. No student has been identified in this thesis. As mentioned earlier, on completion of this present study, other original data files and documents where students are identifiable will be destroyed.

6.3 Data description

Study data and data sources will be described and summarised in this section.

As explained at the beginning of the chapter, data were not received in a consistent format from all sources. The marks data for each student were either their mean IISA module and GSA module marks, or their raw scores for their GSA module IISA and GSA coursework component elements.

The four principal data sources for this present study are designated A, B, C and D in this thesis and in Table13, which lists all 18 data sets and the sample sizes.

Table 13: Study data sets

A Data	An	B Data	Bn	C Data	Cn	D Data	Dn
A1	22	B6	53	C7	35	D11	25
A2	26			C8	52	D12	31
A3	48			C9	38	D13	582
A4	49			C10	52	D14	658
A5	74					D15	751
						D16	285
						D17	445
						D18	844

In the table, the letters A, B, C and D designate the data sources and the numbers 1 to 18 designate the individual data sets. Three of the study data sets (B, C and D) were from the official student marks records from the UK. The data for set B were derived from figure 7 of Knight (2004).

The first year of a first-degree was usually referred to as L1, level one. Similarly, labels L2 and L3 refer to the second and third study years of a first-degree. An integrated Masters first-degree will include a fourth study year (L4). Other, non-integrated Masters post-graduate degrees will be at level four (L4). (Level zero (L0) would be a non-contributing foundation year prior to studying for a first-degree.)

6.3.1 Data from source A

As previously noted, data source A was an Australian HEI. The data comprised of data sets A1 to A5. The data were from the same student cohort studying a PG programme. Their teaching and summative assessment was by the same small group of academics.

In this data set, the data were GSA module IISA and GSA coursework component element marks. The module marks were relatively undiluted by the weight of the IISA contributions. The ratio of IISA and GSA assessment items marks in the modules varied from around 1:1 to 2:1. See section 6.3.4.1 for a discussion on the dilution effect of including IISA and GSA marks in the same total.

As with the other data in this present study, these were from modules studied between 2004 and 2008 inclusive.

6.3.2 Data from source B

Data set B6 was, as was noted earlier, synthesized from figure 7 in Knight (2004:74). This was done by printing out a paper copy of Knight's figure 7 at twice-full size and measuring the position of the data points.

6.3.3 Data from source C

The C data included four years of module marks and three years of degree classifications, labelled C7 to C10. The 2007-8 year cohort had not completed their programme. At the time of the data collection, they had not had a degree classification recommended. The summative assessment marks hierarchy for the C data programme is shown in Figure 7.

The C data set was from a single degree programme taught by one school, in the Science Faculty at a university in the northeast of England. There was no attempt to collect marks for students based outside the teaching school/department. These could have included students studying a Natural Science programme based outside the teaching department, as well as ERASMUS students, i.e. exchange students from other European universities. Potentially, a number of other schools and departments could have been involved. This could have led to the difficulty of negotiating data access for a small number of students.

The only GSA module in this school was common to all undergraduate programmes taught by the school. It was studied at level two (L2). Some students not based there, had the opportunity to study this module as an option at L3 of their programme. Students who studied the module at L3 could not be identified.

Both GSA and IISA module categories marks contributed to the student year mark. This GSA module was double weighted; there were only four IISA modules for that study year. At

assessment component level, separate marks for coursework and a written examination contributed to the GSA module mark. There were two coursework elements. These were the group project individual report marks, and those for the group project itself. One IISA and one GSA coursework component contributed marks to the GSA module coursework. Coursework and written examination component marks contributed to the marks for the other IISA modules (ISA1 to 4). These have been omitted from Figure 7 (the C data, module hierarchy model) for clarity.

In the C data module, the double weighted GSA module contributed 40 points to the 120 points the students needed for their programme year score. The coursework mark contributed 24 points to the module mark. The coursework GSA element contributed 12 points to the coursework mark. The group project mark therefore contributed 10% to the programme year mark.

6.3.4 Data from source D

Data from source D was collected in the form of more than 32000 lines of data in a single spreadsheet. It was from more than 3600 undergraduate subjects. This was 88% of the data collected for this present study. Around 3100 of the data subjects were eligible for inclusion in this study. This was around 79% of the data used in this study.

For the duration of this study, the university in the northeast of England that provided the data provided more than 1300 modules for study at undergraduate level. Thirty-one of them were identified as using GSA and marks data from these modules were included in this study. This compares to, for example 500 group work modules at Loughborough University during a similar period *"representing every department at the institution"* (Loddington 2008:5). The marks weightings of the IISA and GSA modules, components and elements were not identified in the data.

Most of the exclusions were because there was only IISA data for the student. A possible explanation for this GSA data omission was that the data were from students who although they did not study any GSA module themselves, were studying the same programme as those who were. They would then have been included in the results of a data base query of programmes that included GSA modules, and did not exclude students who did not study them. This possible explanation was unverifiable. The IISA only data were excluded from the present study.

Data from source D was from three study years. Data from source C was from four study years.

The first three years of data from source C are also included in the D data. These are module marks for around 125 students. In the data from source D, around 3% were students from the C data source. They were double counted in this study. There are implications for comparisons and conclusions because of this. They need to be interpreted with this double count in mind. Some of the data from source C was anonymous within it, and was therefore inseparable from it. Avoiding this by omitting the module code from the list of GSA modules included with the data request would have meant that data from source D would have been incomplete. This would have raised issues surrounding how to deal with this omission.

During the data collection phase of the present study, a module was found that allowed the students to choose the method of their assessment. Students could choose between an IISA mark and a GSA mark. The rationale for this was that it would reduce the risk of would-be none-contributors affecting the marks of the other group members. Having some groups of students select their assessment method bypasses several points of the rationale for using GSA, (see section 3.5). Because of this inconsistency, data from this module has not been included.

6.3.4.1 The dilution effect of IISA marks

In the data from source D, the marks were for the whole module and they included contribution from both IISA and GSA assessment items. There was no separation of contributing marks. Modules were categorized as either IISA or GSA. Those categorized as GSA modules were from marks contributions from both IISA and GSA items. The methods of deriving the marks within the modules, and their contribution weightings, were unknown.

The number of assessment items contributing to each of the module marks in the data from source D is unknown. It was known only that the anonymous programme from a known faculty and year included one or more IISA items and one or more GSA items. There were no modules found that only assessed GSA items. From this present *study* point of view, this would have the effect of diluting any contrast between marks from IISA and GSA modules. The extent of the dilution was unknown.

It is clear that for some of the GSA modules, the actual GSA item will have only contributed a small percentage to the student year mark. It would depend on the number, weighting and combination of categories of assessment summative assessment items in the module. Typically, from two to four items were assessed in GSA modules, although there were some beyond this range. In each

module, only one of these was GSA. In addition, some programmes included more than one GSA module.

6.4 Data description and collection chapter summary

This chapter has introduced the data and outlined its sources. Why undergraduate data were preferred for this present study was explored in section 6.1. Ethical considerations were discussed in section 6.2. The treatment of potential outliers was explored in section 7.1. It also explained why they were all included. The data and the four data sources were detailed in section 6.3. The dilution effect that IISA item marks have on GSA module marks was discussed in section 6.3.4.1.

The next chapter, 7, presents the different data analysis methods and the results of this present study.

Chapter 7. Data processing, analysis and results

The previous chapter focused on data collection whereas this chapter relates to the data processing, analysis and results.

From section 1.1 (Table 1), the alternative hypothesis H_1 is that the two summative assessment methods, IISA and GSA, each have a different impact on students overall marks. This would lead to different degree classification awards for students, depending on the mix of IISA and GSA marking methods used for the programme modules. In particular, it could most affect those whose abilities are towards the ends of the individual ability scale.

Several different methods of data analysis were used to explore the differences between IISA and GSA marks data. There are eight sections in this chapter, including a summary section. Section 7.2 further describes and presents summary statistics for the individual data sets from the four sources. It presents distributions, as well as means and standard deviations. Section 7.3 presents analysis of the correlation between IISA and GSA marks within and between data sources. It is the key relationship between the data in this present study. This includes the relationship between the single module data; set C, group and individual element items from assessment components. Section 7.4 explores regression as an extension of the correlation. Section 7.5 used Multilevel modelling as an alternative way to look at the data. It is an extension of regression analysis. Section 7.6 applies Meta-analysis and Forest Plots to present an exploration of data set sampling variations around a single relationship. Section 7.7 is simulation. It seeks to simulate the data by making simple assumptions. Section 7.8 is a chapter summary.

7.1 Treatment of potential outliers

This present study has explored the relationship between specific pairs of assessment category data (i.e. students IISA and GSA marks), e.g. see section 7.4. Where, for example, a student scored zero for one or other of IISA or GSA then this was potentially of substantive importance to the study. For this reason, all potential outliers have been included in all the data sets. While almost all marks of zero are due to non-submission, some other reasons for their exceptional marks may have been GSA issues. For example, the student may not have been able to prepare properly for an assessment. They may not have engaged with it. On the other hand, they may have engaged with it at the expense of their other modules and assessment items. There may have been a combination of a whole gamut of reasons for their exceptional performance in one or more assessment. Two examples of possible outliers in the study data are shown in Figure 10.

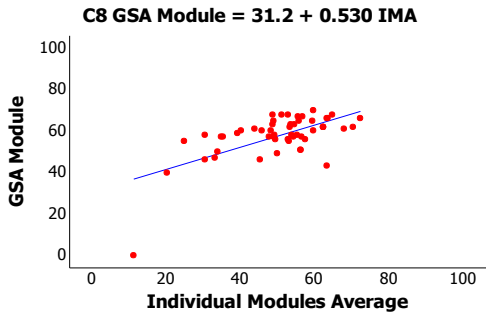


Figure 10: Example of possible outliers

The data are from subset C8. (All the charts are shown in Appendix 14 and Appendix 12.) The panel shows two possible outliers. One is from a student in the high IISA ability category, towards the right of the chart. He/she had a relatively poor GSA mark, (43), compared to the regression prediction, and had scored 63 for their IISA assessment item. This student plainly had a problem with the GSA item despite, or perhaps even because of, their relatively high IISA score. The other student scored zero for their GSA module. In this latter example, it seems clear that since the student also scored the lowest IISA marks, they did not benefit from their studies of this module at least, to the extent that they could have.

7.2 Data distributions and summary statistics, from four data sources

Analysis of the data from the four individual sources is presented below in sections 7.2.1 to 7.2.4. The data are summarised and compared in section 7.2.5.

7.2.1 Source A data analysis

The initial description of data source A is in section 6.3.1. There were 219 data subjects in total. Frequency histograms of marks distribution are presented. This part of the analysis combined the data from all five data sets because there were relatively few subjects in the individual sets and to show that overall, the data supported the Downie (2001) finding explained in section 3.7. Figure 11 is a marks frequency chart of the IISA and GSA marks study data from the combined data sets A1 to A5.

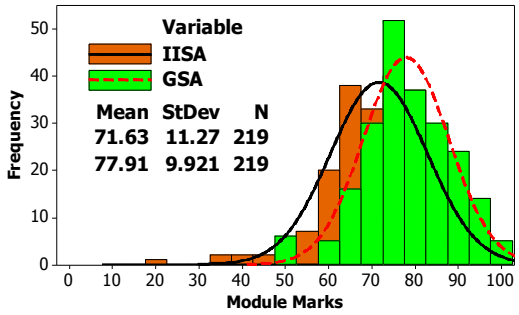


Figure 11: Marks frequency histogram of data source A distribution

The charts have normal curves fitted. The individual module sample sizes ranged from 22 to 79 (see Table14). A visual inspection of the distribution chart Figure 11 shows that the shape of both categories of marks distributions approximates normality. In addition, it shows that the top of the range of combined marks in these modules was towards the high end of the 0 to 100 x-axis module assessment marks scale, almost hitting a ceiling. The higher GSA mean value with a lower standard deviation is typical of this type of data, see Downie (2001:7) and section 3.7. It is also typical of the relative distribution of the other IISA and GSA data in this study. The collective GSA marks had a mean of 77.9 and a standard deviation of 9.9. The IISA marks had a mean of 71.6 and a standard deviation of 11.3, this difference supported the Downie finding, mentioned in the opening lines of this section. On the other hand, it is apparent in Figure 12 to Figure 16 that this higher GSA mean and smaller SD distribution is not typical in all the individual data sets from source A, e.g. data sets A1 and A2, (also, see summary chart in section 7.2.5).

This pattern of a higher mean with a lower standard deviation is not exclusive to GSA. It has been found to be a feature of coursework assessment generally. Simonite's study was not specifically exploring GSA in her 2003 work. Her study focussed on the effect of *IISA* assessment (i.e. exam marks); nevertheless, the parallels between her study and this one are clear. She commented that:

> *"As expected, in modules in which all or part of the assessment was based on coursework, students achieved significantly higher mean marks, other things being equal, than in modules using only examination assessment".*

(Simonite 2003a:463)

Figure 12 to Figure16 show marks frequency histograms of the separate IISA and GSA data sets; in addition, they have normal curves fitted.

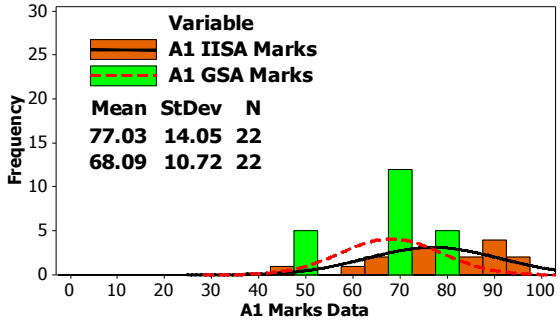


Figure 12: Marks frequency histograms of A1 IISA and GSA data

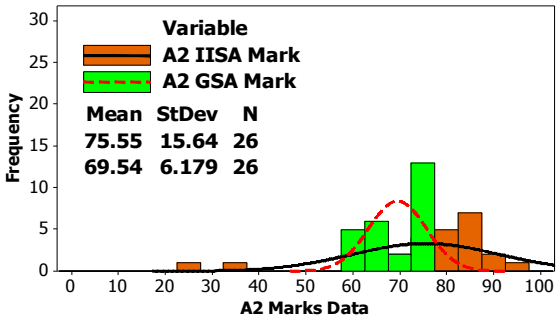


Figure 13: Marks frequency histograms of A2 IISA and GSA data

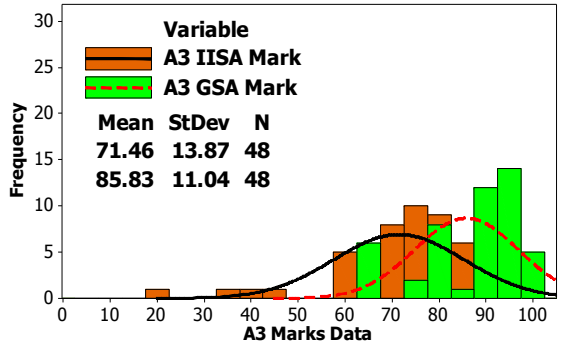


Figure 14: Marks frequency histograms of A3 IISA and GSA data

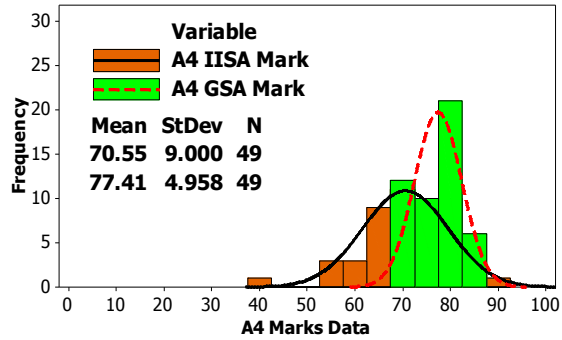


Figure 15: Marks frequency histograms of A4 IISA and GSA data

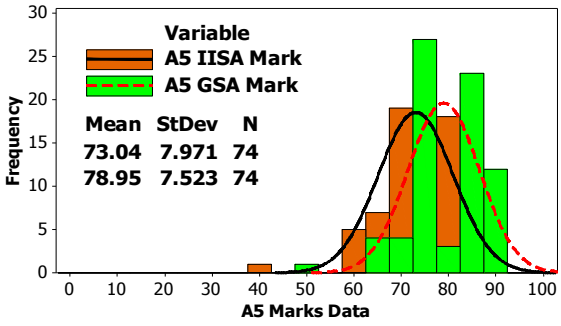


Figure 16: Marks frequency histograms of A5 IISA and GSA data

All of the data sets are approximately normally distributed, bearing in mind the small numbers involved in some cases, although some of the data shows a ceiling effect with some of the charts

showing the data skewed towards the right, the high end of the zero to 100 marks range. The A3 GSA marks data histogram (Figure 14) shows this data ceiling effect most clearly. On the other hand, the data distributions are approximately normal bearing in mind the low numbers of data cases in some of the A data sets.

The IISA and GSA marks data were tested for statistical significance. A t-test is parametric. Parametric tests rely on normally distributed data but the data were not normally distributed, they only approached normality. However, Glass et al. (1972) established that the t-test was also robust against the violation of assumptions of normality. This, and the data having a distribution approaching normality, meant that a t-test is a valid test of statistical significance between IISA and GSA pairs for this data. The sample sizes, means, standard deviations and results of the IISA and GSA marks statistical significance tests are in Table 14.

Table 14: Data source A summary statistics

		IISA		GSA		Paired t-test
Data Set	N (student cohort size)	Mean Mark	Standard Deviation	Mean Mark	Standard Deviation	p (df)
A1	22	77.0	14.1	68.1	10.7	0.030 (21)*
A2	26	75.5	15.6	69.5	6.2	0.066 (25)
A3	48	71.5	13.9	85.8	11.0	0.000 (47)**
A4	49	70.6	9.0	77.4	5.0	0.000 (48)**
A5	74	73.0	8.0	79.0	7.5	0.000 (73)**

*p < .05, **p < .01, - = not significant

Four out of the five A data sets showed differences in mean scores that were statistically significant. The t-tests showed that the A3, A4 and A5 data were statistically significant at the 99% level ($p < .01$). The A1 data were statistically significant at the $p < .05$ level. The t-test resulted in a higher p-value of 0.03. The A2 data was not statistically significant. The p value was 0.07. A one-way ANOVA showed that the means of the IISA marks were not statistically significantly different, i.e. they could have been from the same population, and that the mean values of the GSA data did vary significantly ($p < .05$). This is illustrated in Figure 17, the cohort IISA marks means and confidence intervals, (and in Figure 18).

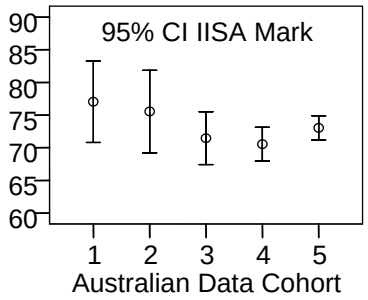


Figure 17: Data source A cohorts IISA marks means and confidence intervals

All the IISA confidence intervals overlap. There is no statistically significant difference between the cohort IISA data, which means that they could have been drawn from the same group. The confidence intervals for cohorts 1 and 2 are greater than for cohorts 3, 4 and 5. Cohorts 1 and 2 have the smallest sample sizes. Cohort GSA marks means and confidence intervals are shown in Figure18.

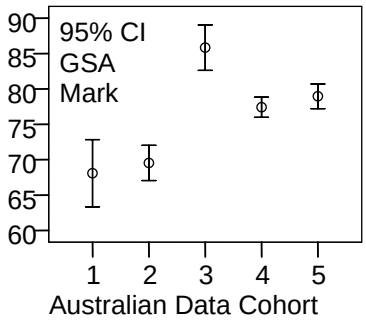


Figure 18: Data source A cohorts GSA marks means and confidence intervals

Only some of the confidence intervals of the cohorts GSA marks overlap, i.e. are not significantly different. The GSA confidence intervals are shorter than the IISA confidence intervals, despite the sample sizes being equal. Cohort A1 data has the largest confidence interval. As shown in Table14, the A1 data set had the lowest number of data subjects (22).

The A1 and A2 mean GSA marks are lower than their mean IISA marks, confounding the Downie (2001) finding, mentioned earlier. The pattern of the mean IISA and GSA marks in sets A3, A4 and A5 are however typical. In these, the GSA mean marks are higher than the IISA marks to which they are paired. On the other hand, all of the GSA standard deviations, being smaller than those of their paired IISA SDs, are also typical.

7.2.2 Source B data analysis

Figure 19 is a histogram of source B data.

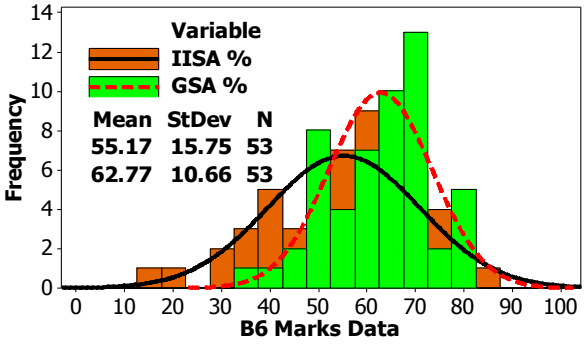


Figure 19: Marks frequency histogram of data from source B

This histogram shows a similar distribution to that of the total data in data source A, but without the

ceiling effect. The distribution seems to approach normality. It also shows the typical higher GSA mean and lower standard deviation differences that were noted by Downie (2001) and previously mentioned in section 3.7. The GSA mean mark was 63.8 and the GSA SD was 10.7. This compared to the IISA marks where the mean mark was 55.2 and its SD was 15.8. The means were statistically significant ($p < .01$). The B data are summarized in Table15.

Table 15: Data source B summary statistics

		IISA		GSA		Paired t-test
Data Set	N (student cohort size)	Mean Mark	Standard Deviation	Mean Mark	Standard Deviation	p (df)
B5	53	55	16	63	11	0.002 (52)**

**p < .01

7.2.3 Source C data analysis

Figure 20 shows a histogram of the collective data for set C modules IISA and GSA element assessment item marks.

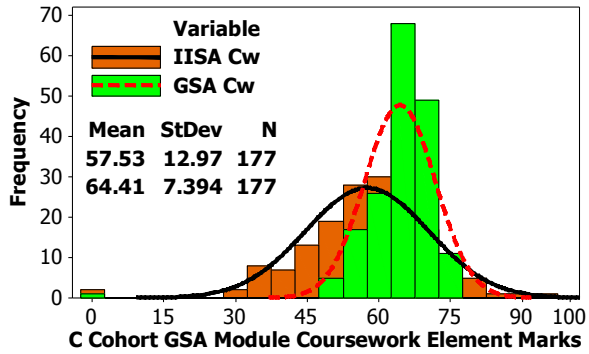


Figure 20: Marks frequency histogram of the combined data from source C element marks

In Figure 20, all four of the study years data, which were received separately, have been combined. The figure charts the distributions of the two marks categories and illustrates differences in their means and standard deviations. In this example, the difference between the two marks categories was the most pronounced of all the data. This is because the data are compared at the module element level, the lowest level of summative assessment marking (see Figure 7). At this level, the GSA module coursework group project mark can be compared directly with the IISA mark from the same coursework. They are not module marks from a higher level, so the GSA item marks are not diluted with marks from additional IISA items. The histogram shows that the frequency distribution of both marks categories from the collective C data approaches normality. The small cluster around zero seems likely to be typical of students module marks data. It is also interesting to note that twice the number of students scored zero for their individual element than for their group element. From this, it could be hypothesised that GSA is more supportive of lower individually attaining students than IISA.

The IISA marks have the lower mean mark (57.5) and the higher standard deviation (12.9). The GSA marks have a higher mean and a lower standard deviation (m = 64.4, SD = 7.4) compared to the IISA marks. In this collective C data set, the mean mark plus and minus three standard deviations are inside the 0 to 100 marks range. The ceiling problem that affected data sets A1 and A2 is not exhibited in the C data.

Table 16 summarises the data from source C.

Table 16: Data source C summary statistics

		IISA		GSA		Paired t-test
Data Set	N (student cohort size)	Mean Mark	Standard Deviation	Mean Mark	Standard Deviation	p (df)
C7	35	62	8.0	68	2.9	0.000 (34)**
C8	52	57	10	66	4.7	0.000 (51)**
C9	38	60	17	65	4.0	0.115 (38)
C10	51	54	12	61	7.1	0.000 (50)**

*p < .05, **p < .01

All four years of data sets support the Downie (2001:7) finding regarding the higher GSA means and lower standard deviations compared to individual assessment marks, also, see section 3.7). The GSA mean marks were all higher than the IISA mean marks and additionally, the GSA standard deviations were all smaller than the standard deviations of the IISA marks.

The t-tests showed that the differences between the IISA and GSA marks of three out of the four of data sets, i.e. C7, C8 and C10, were statistically significant at the 99% level, indicating that they were from different year groups. The difference in the C9 IISA and GSA marks data was not statistically significant. A one-way ANOVA showed that the mean values of both categories of marks varied significantly.

7.2.4 Source D data analysis

Figure 21 shows a histogram of the D data module marks frequencies for the IISA and GSA modules and for all the module marks combined.

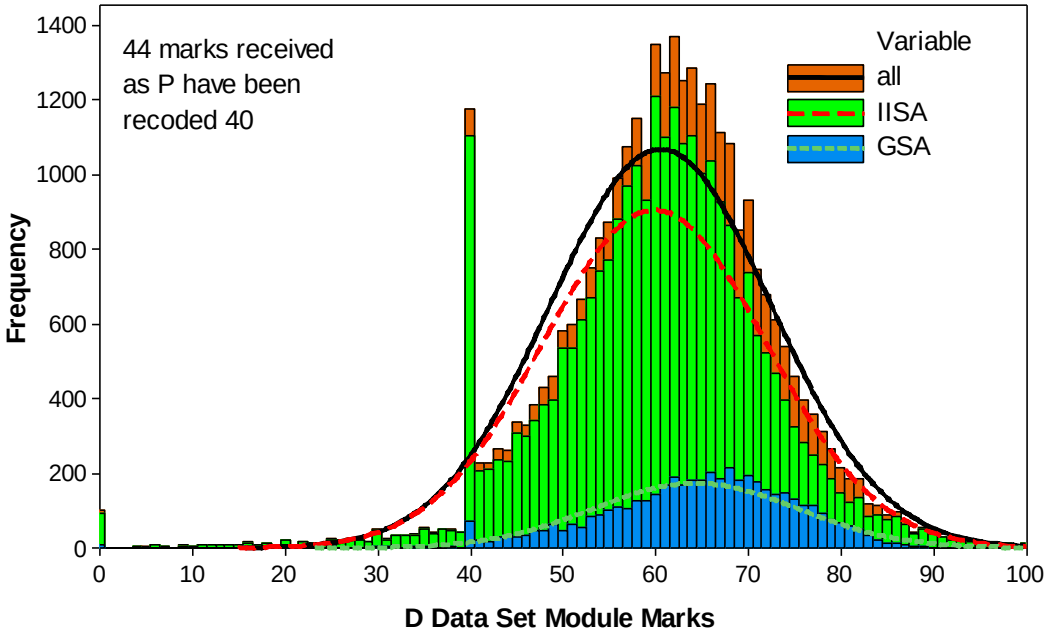


Figure 21: Marks frequency histogram of data source D module marks distribution

The means, standard deviations and the sample sizes in the whole study are shown in table 17.

Table 17: Means, standard deviations and the number of data subjects from Figure 21

Variable	Mean	Standard Deviation	N (sample size)
all	60.58	12.02	32158
IISA marks	59.92	12.06	27353
GSA marks	64.36	11.05	4805

Normal curves have been included in the distribution chart. The frequencies of the two categories of marks data are approximately normally distributed, as is the distribution of the combined data (labelled all in the chart). As the data were at module level, the numbers of IISA data cases far outweighed those of GSA data. In addition, like data source C, and unlike data source A, the distribution of this collective data did not show any ceiling effect or skew.

The histogram shows the typical higher GSA mean (64.4) and lower GSA standard deviation (11). This GSA SD varies only slightly from that of the IISA data compared to the other data sources used in this study. The standard deviation of the IISA marks was 12.1. The IISA marks mean was 59.9. The values of the differences (mean 4.4, SD 1.01) were, on average, less than in the other study data. This was especially true of the difference in standard deviations. The comparatively small difference seems likely to be a result of the dilution effect caused by IISA assessment item marks being included in the module mark. It could also be the result of the much smaller sample size compared to the IISA data (4805 cf. 27353).

In Figure 21, the superimposed normal curves all appear to be slightly to the left of the marks distributions. This is because of the tails on the left-hand side of the data, towards the zero point. This is the result of the combined effect of the 40 and zero marks peaks. The marks frequency spike at the 40-module mark point on the x-axis is in the order of 3.5% of the data. Between them, 631 data source D students received 1133 marks of 40% for at least one of their modules. Undoubtedly, some students would have received this score even if the pass mark had been different but there are other possibilities set out in the next paragraph. One student was awarded a score of 40 for 11 of their qualifying modules. They, unmistakably, had problems either with or during their studies.

One of the reasons for the frequency spike at the 40 marks point could be that it included marks from successful module re-sit attempts. Re-sit attempts for D data students at this university in the northeast of England were not allowed at the final year assessments. Where they were successful, the award was only 40 marks, a bare pass. This was to discourage students using deliberate failure as a strategy to allow them to re-sit the assessment. The additional study time would give them an unfair advantage over their peers, and allow them a second chance to improve their degree classification. The second reason is that it could include marks resulting from marker or exam board discretion.

A senior university academic also confirmed the 40% spike trend in a personal communication. She/he had a teaching and assessment role at the same university in the northeast of England. On seeing a copy of Figure 21, she/he confirmed that it was entirely consistent with assessment practice. She/he explained that the number of re-sits in their school or department was consistent with the study data 40-mark spike frequency.

There was also a much smaller spike at the zero marks point. At this HEI, an unauthorised late submission of assessed work was not marked. It received a mark of zero. This may account for at least some of the frequency of zero marks. The reason for the step in the histogram immediately below the 40-module mark is unknown but it could be because markers pay more attention to borderline marks. Exam Boards also could make a different award after reviewing the candidate's university record.

Other reasons could include those resulting from an appeal by the student. The appeal could be, for example, on medical or equal opportunity grounds. These reasons could however affect the marks awarded at any level, not just those near to 40. In Figure 21 there does seem to be larger steps in the data before the marks break points between the 2.2, 2.1 and First degree classification levels. This aspect of student assessment marking was not the focus of this study.

In addition, 44 subject cases out of more than 32000, i.e. < 0.14 percent, had been given a *P* threshold module mark. No student had more than one *P* threshold module mark recorded in the marks data. For this present study, a mark of 40 replaced the letter *P*. It seems unlikely that this small number of cases would have any significant statistical or educationally substantive effect on the results. The alternative to not recoding would have been to exclude from the study those students whose marks included a *P* designation. This account would then have been incomplete. A summary of the D data is shown in Table 18.

Table 18: Data source D summary statistics

		IISA		GSA		Paired t-test
Data Set	N (student cohort size)	Mean Mark	Standard Deviation	Mean Mark	Standard Deviation	p (df)
D11	25	63	3.9	66	4.7	0.000 (24)**
D12	31	63	5.0	65	4.1	0.001 (30)**
D13	582	61	11.1	65	11.3	0.000 (581)**
D14	658	61	11.1	66	10.0	0.000 (657)**
D15	751	60	11.5	64	10.9	0.000 (750)**
D16	283	58	9.5	64	8.8	0.000 (282)**
D17	445	58	8.4	63	9.2	0.000 (444)**
D18	844	57	9.8	63	9.4	0.000 (843)**

**p < .01

All eight D data sets support the Downie (2001:7) finding that GSA mean marks were higher than IISA mean marks (see section 3.7). Three of the GSA standard deviations, i.e. D11, D13 and D17, were however larger than their paired IISA SD, confounding Downie's SD relationship finding. All the data set means were statistically significant (p < .01) at the 99% level, i.e. the samples were all different to one another. A one-way ANOVA showed that the means of both categories of marks varied significantly at the 99% level.

7.2.5 Data Analysis summary

This section summarises the study data sets. (See Appendix 15 for summary statistics of the eighteen data sets.)

The mean IISA mark for each data set ranged from 50.2 for data set C8, the single Science Faculty module from a university in the northeast of England, to 77.0 for data set A1, an Australian PG data

set. The overall mean IISA mark between the data sets was 62.4. The range of the mean GSA marks for each data set was from 57.5 for data set C10 (the single Science Faculty module), to 85.8 for data set A3 (Australian PG data). The overall mean GSA mark was 66.2. The overall difference between the mean data set GSA mark, was 3.8 (66.2 - 62.4). The difference between the mean combined data marks, where n, (the number of students, not the number of groups of students,) is 4069, was 4.8 (64.8 – 60). The difference between the mean IISA and the mean GSA mark between the data sets ranged from -8.9 for data set A1 to 14.4 for data set A3.

The IISA standard deviations varied between 3.9 for data set D11, an Arts and Humanities module, and 15.8 for data set B6. The overall mean IISA SD was 11.1. The GSA standard deviations varied between 4.1 for data set D12 (an Arts and Humanities module), and 11 for data set D13 (Science). The overall mean was 8.4. The differences between the mean IISA and the mean GSA SD was from -0.80 for data set D11, to 9.5 for data set A2. The mean difference between the standard deviation of the GSA marks and the SD of the IISA marks was 2.7 (11.1 cf. 8.4). The mean IISA and GSA mark of the 18 cohorts are shown in Figure 22.

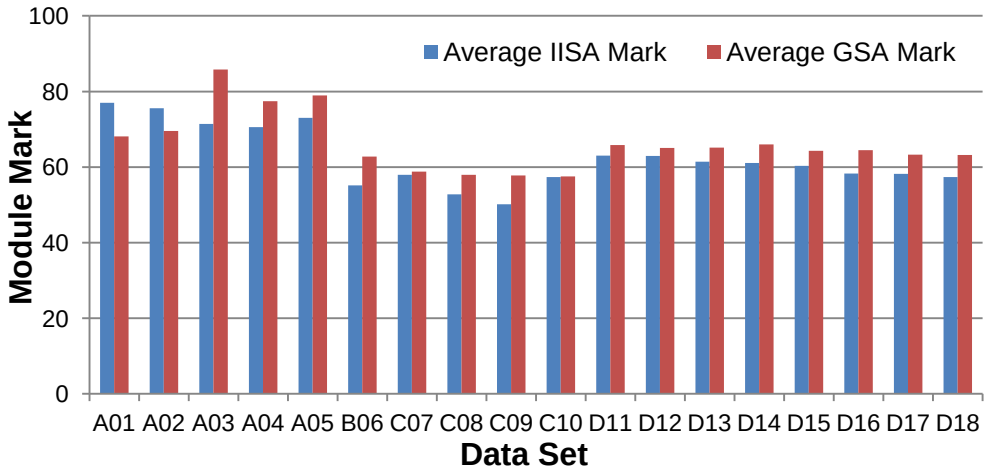


Figure 22: IISA and GSA mean marks

The bar chart shows the mean IISA and mean GSA marks for each Data set. The value and direction of the marks difference is the advantage or otherwise for an *average* student in that cohort. The chart shows GSA marking to disadvantage the average student in only two cohorts, A1 and A2. Additionally, the marks of both categories in data sets A1 to A5 are noticeably high, compared to their equivalent data points in the other sets.

7.3 Correlation

In this section, alternative hypothesis H_3 will be explored. As mentioned elsewhere (sections 4.6.2. and 5.6), the correlation between the IISA and GSA marks is a key relationship in this present study. A low correlation between IISA and GSA marks is an indication that the marks are measuring two different constructs. This part of the exploration used the Spearman's rho correlation because it is a non-parametric variable and the data distributions were not perfectly normal. In practice, there was very little difference between the results for Spearman's rho correlations and Pearson's r, which assumes that the data is normally distributed.

Data from source C were the only data where sufficient detail had been collected to allow this level of analysis. (The summative assessment structure of the single Science Faculty programme where the C data was collected is shown in Figure 7.) Seven sub-totals and levels of assessment marks were either collected in, or could be derived from, the C data. These were marks for the:

Year
GSA module
Calculated average of the students IISA modules
GSA module GSA coursework
GSA module written examination (IISA)
GSA module coursework individual (IISA) report element
GSA module coursework group project (GSA)

Only one of these pairs was of interest to this study and was explored. This was the pairing of the last two data pairs in the list, the GSA module coursework individual (IISA) report and the group project (GSA) elements. Some of the others would add little to an understanding of the impact of GSA marking. For example, the correlation between the GSA module IISA item written exam, and the mean of the other IISA module marks, may not be especially useful. They are both IISA items. There would also be a smoothing effect from averaging the IISA module marks. The written examination mark would also depend on the student's effort and ability covarying with their learning from the group project. In a different study, this might contribute to knowledge of its robustness by triangulating data findings.

There are two parts to this correlation section.

Section, 7.3.1 presents the correlation coefficient categories used in this study.

Section 7.3.2 presents and discusses the correlations between the 18 data sets.

7.3.1 Correlation coefficient categories

For this study, in order to facilitate the discussion, correlations were put into three categories: Low: <= 0.33, Medium: 0.34 to 0.66, and High: 0.67 to 1, (see Table 22).

Table 19: Correlation categories

Low	=< 0.33
Medium	0.34 to 0.66
High	0.67 to 1.00

7.3.2 Correlation between IISA and GSA marks in 18 data sets

Figure 23 is a chart of the Spearman's rho correlations between the IISA and GSA marks of the GSA modules coursework marks, and/or of the programme modules marks for each of the 18 study data sets.

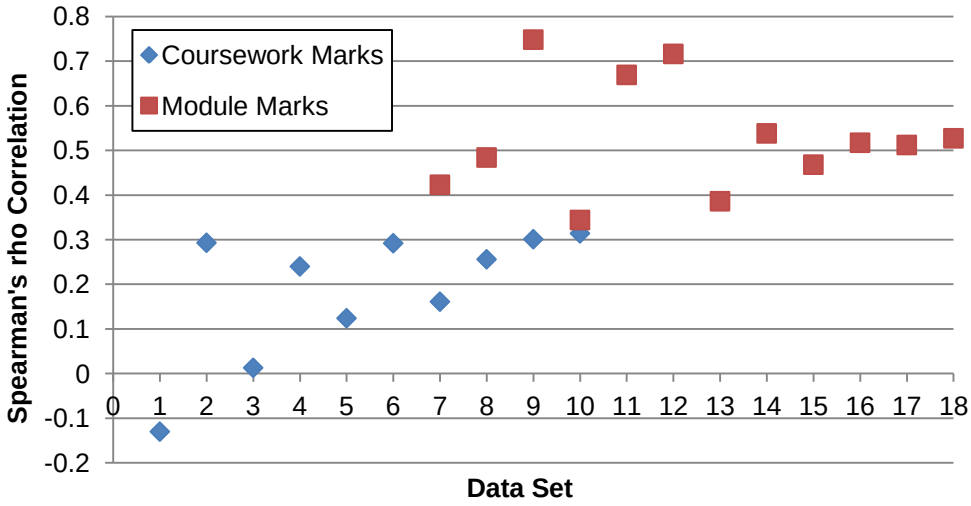


Figure 23: 18 Data sets GSA and IISA Spearman's Rho correlations chart

The correlations data is shown in Table 20.

Table 20: Correlations between IISA and GSA data using Spearman's Rho

Data Set	Sample Size (N)	Correlation (rho) between the IISA and GSA coursework element marks	Programme module marks correlation (rho)***
A01	22	-0.13	
A02	26	0.29	
A03	48	0.01	
A04	49	0.24	
A05	74	0.12	
B06	53	0.29*	
C07	35	0.16	0.42**
C08	52	0.26	0.48**
C09	38	0.30	0.75**
C10	52	0.31*	0.34**
D11	25		0.67**
D12	31		0.72**
D13	582		0.39**
D14	658		0.54**
D15	751		0.47**
D16	285		0.52**
D17	445		0.51**
D18	844		0.53**
Rho mean		0.18	0.53
N mean	226		
A mean		0.11	0.50
C mean		0.26	0.56
D mean			0.54
2-tailed significance * p < .05 ** p < .01 *** Correlation between the GSA module marks and the remaining IISA modules marks confirms the prediction that these correlations would be higher due to the dilution effect (see section 6.3.4.1).			

All the data sets IISA/GSA correlations between *elements,* at the *coursework* level, were low (<0.34). They ranged from -0.13 to 0.31. The mean correlation was 0.18. In addition, only the C10 and B6 data were significant, and only at the 95% level ($p < .05$). On the other hand, all the *module marks* data sets have either a medium or a high correlation. They were all significant ($p < .01$) and ranged from 0.34 to 0.75, and the mean was 0.53. The highest correlation was between the data set C9 IISA and GSA marks. At the *module marks* level, the GSA marks data, and the correlation, were diluted with IISA assessment item marks, as mentioned earlier in section 6.3.4.1. This supports the hypothesis that, overall, IISA and GSA were measuring different constructs or that one or both measures were unreliable. Hypothesis H_3 (see Table 1) was not disproved.

7.4 Regression

The previous section analysed correlation between the data types. As indicated elsewhere, (e.g. section 5.6,) simple regression was the first analysis applied to early data in the run-up to this study. This method is quick to apply, widely used and produces results that are readily interpreted. On the other hand, simple regression ignores the hierarchical nature of educational data.

This regression analysis section is in four parts. These are 7.4.1, which explains the importance of the two types of regression chart to this study, 7.4.2, single-line regression, 7.4.3, dual-line regression, including section 7.4.3.1, which explains the privileging effect on dual-line charts. Section 7.4.4, is a results section.

7.4.1 Two types of regression chart

There were two types of regression chart used in this study. The first was a single-line IISA/GSA data scatterplot. This type of regression chart is very common among researchers. It illustrates the focus of this study, i.e. the regression slope and the correlation between the data categories. An example is shown in Figure 24.

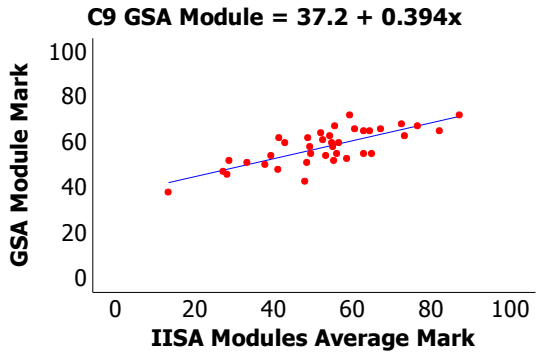


Figure 24: Single regression line chart

An example of the second type of regression chart is shown in Figure 25 and shows a dual-line regression chart of the same data of the type used in Almond (2009).

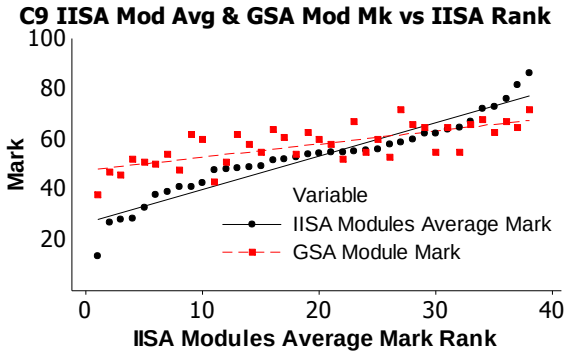


Figure 25: Dual regression lines chart

It shows the relationship between the data categories against students IISA modules mean mark. An annotated generic dual-line chart is shown in Figure 26.

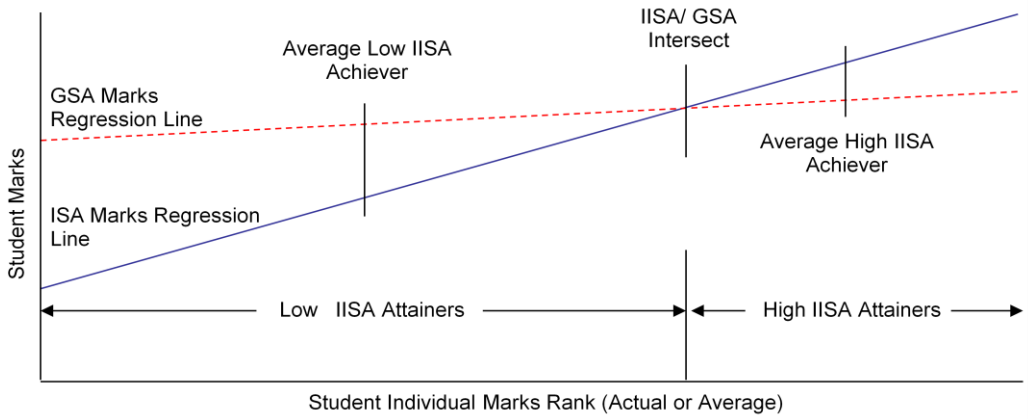


Figure 26: Annotated generic dual-line IISA and GSA chart

NB: In this study, in dual-line charts similar to Figure 26, the GSA mark is either at the module level or at the coursework component or component element level (see Figure 7).

In Figure 25, the chart shows the IISA and GSA data from data set C9. This data set was selected for clarity in illustrating the two types of regression chart for two reasons. Firstly, it had the highest overall module correlation (0.75) so the slope of the regression line would be the most pronounced. Secondly, the C9 data set was one of the lowest sample sizes overall and the lowest C data sample size (38) so the lines on the charts are less obscured by the data points, the spread of which is more easily discernible.

The x-axis of the single-line chart in Figure 24 shows students mean marks from all of their IISA modules. Each data point chart represents one student. The y-axis shows their GSA module mark. The slope of the single line indicates how a change in the value of the mean of the students IISA marks predicts a change in the value of their GSA marks for this student cohort. The flatter the line, the less the predicted impact and the less the two categories could be assumed to be assessing the same student attributes. In this example, there is a positive relationship between the marks from the two categories. The slope of the regression line is 0.39 and the intercept is 37.2. In addition, the grouping of the data points in the single line chart illustrates their spread, more easily enabling potential outliers to be identified for further investigation. The relative positions of these data points suggest that there were no outliers from this student cohort. The scales of both axes of single line charts are consistent throughout the study data sets. From this, they can be easily compared between the study modules (see for example Appendix 14, Appendix 12 and Appendix 15).

For this study, the point which distinguishes IISA high and low achiever marks, has been taken as the point where x = y. It occurs at some point along the x-axis but it cannot be determined by

observation of a single-line chart. This point is more easily seen and more precisely illustrated graphically on a dual-line chart, where the two regression lines bisect. The regression lines of sixteen of the eighteen dual-line scatter plots bisect within the range of the x-axis, (see Appendix 12). Most of those were between the median point and the highest ranked IISA student mark, i.e. towards the right-hand end of the x-axis. This is explored further in section 7.4.3. Results from the dual-line regression analysis are in section 7.4.4.

A further advantage of the dual-line chart is that the ratio of the advantage, or disadvantage, of one category of data, i.e. IISA or GSA, over the other is more immediately apparent from the relative distances between the low and high ends of the IISA and GSA lines. The spread of the data is however, lost. In addition, there are three disadvantages. First, the scale of the dual-line x-axis varies depending on the number of students in the cohort so it is more difficult to detect possible outliers by observation. Second, it is not as easy to compare the slopes of the regression lines between data sets (in Appendix 12) by observation. Thirdly, it is a non-standard method.

As reported in Almond (2009) and shown in Figure 2, the dual-line chart was the first step in exploring the relationship between IISA and GSA marks. The dual-line chart shows the relationship, i.e. the degree of parallelism, between the bisecting regression lines. In this model, the more closely related the regression lines, in particular their slopes, the greater the likelihood that the two assessment categories were assessing similar attainment measures.

7.4.2 Single-line regression charts

As mentioned above, the variable of interest in the single-line regression chart is the slope of the line, rather than its intercept. The slope of the regression line relates directly to the correlation.

Table 23 listed the correlations between the data categories using Spearman's rho. Figure 27 shows two examples of single-line charts (the full range of which are shown in Appendix 14).

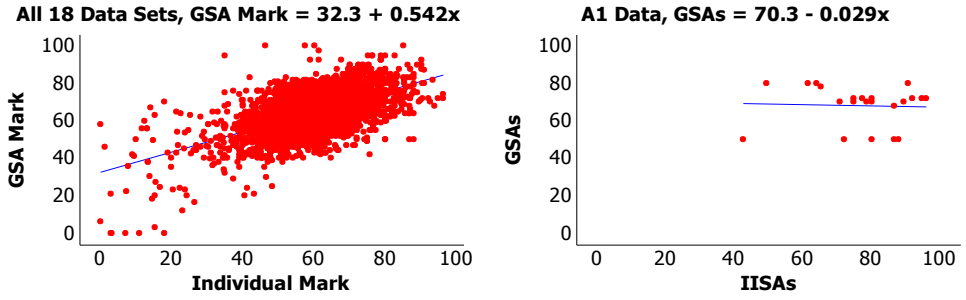


Figure 27: Single-line regression chart examples

The chart on the left shows all of the data points from all 18 data sets on one chart (n, the number of students in the study sample = 4069). The chart on the right shows the data set A1 scatter plot (n = 22). The regression lines on several of the individual charts are quite flat corresponding to the low correlations noted earlier.

The slope of the line confirms the strength or otherwise, and direction, of the relationship between the IISA and GSA marks. The greater the slope, the stronger the relationship between the individual and the group marks. The steepest slope, 0.92, was for data set D11, which had a confidence interval of ±0.33 together with a low sample size, 25. The slope of the combined data was 0.54, with the smallest confidence interval of ±0.012. On the other hand, data set A1 has the shallowest slope of -0.029, the smallest sample size, 22 and, as is usually associated with a small sample size, the largest confidence interval (±0.36). It was also the only negative slope, albeit negligibly so, in the data. Data set D18 has a slope of 0.56, the largest sample size (844), and, unsurprisingly, has one of the smallest CIs, ±0.053.

The next section explores dual-line charts.

7.4.3 Dual-line regression charts

In addition to the correlation between the two data categories, in this study, there were two further points of interest on the dual-line regression charts. The first was the slopes of the IISA and GSA regression lines. The focus was on a comparison between the slopes, and the differences between the y values at either end of the x-axis. The differences in slope indicated the extent that similar student attributes were being assessed. The differences in the values of the marks at either end of the x-axis indicated the ratio of the advantaging or disadvantaging of the GSA marks over the IISA marks. The second point of interest was the bisection along the x-axis. This indicated the ratio of students who could be considered as winners, to those who were disadvantaged by GSA in the data set, as is presented in the regression results section later.

Two different examples of dual regression charts, shown with and without data points, are shown in Figure 28.

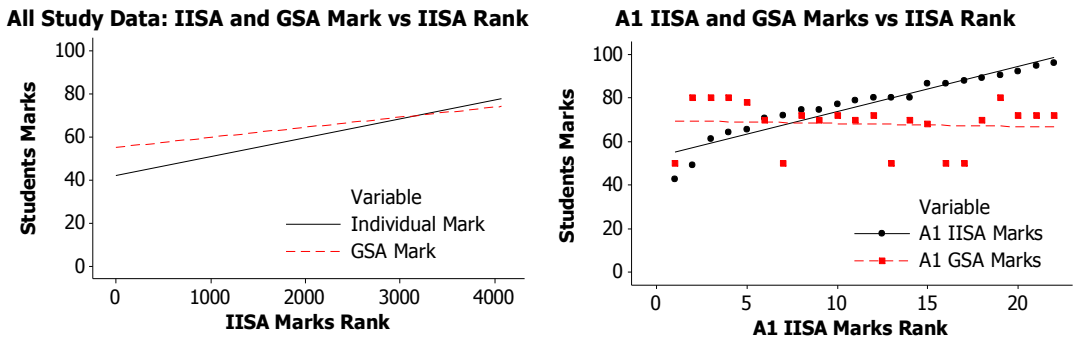


Figure 28: Dual-line regression charts

The dual-line chart in the left-hand panel, without data points, shows regression lines for the collective study data (n = 4069). The right-hand chart, with data points, shows regression lines from the data set A1 students (n = 22). (All of the dual-line regression charts for the 18 data sets and the combined data are in Appendix 12.)

The data from the four different sources were at varying levels of detail. Because of this, the data used in this study was from three different formats. The x-axes of the charts in Figure 28 (and Appendix 12), are the ascending rank order of either; the student GSA module IISA mark, the GSA module IISA items average marks, or the mean of their IISA modules marks.

Typically, the dual-line charts show the IISA and GSA regression lines as bisecting within the x-axis range. As previously mentioned in section 7.4.1, this is predominantly towards the right of the x-axis, towards the higher IISA ranking. This means that generally, the advantage that GSA marking gives the lower ability students is greater than the disadvantage suffered by the high achievers. It also means that fewer students were disadvantaged than were advantaged. Overall, although there are some small differences, each of the eighteen dual-line charts of the study data sets shows the same distribution pattern as the single cohort data in Figure 2.

In data sets A1 and A2, from the Australian post-graduate modules, the IISA and GSA regression lines bisected towards the left hand of the x-axis, i.e. towards the lowest IISA ranking. This meant that more high individually achieving students were disadvantaged by GSA marking and fewer low individually achieving students were advantaged by it.

There were also two other data sets where the regression lines bisection point was exceptional. These regression lines would, if extended, bisect beyond the right-hand end of the x-axis. They

were data sets D11 and D17. This was an important consideration in the meta-analysis, in section 7.6 because it meant that all of the students in these two cohorts were advantaged by the GSA marking. In these cohorts, there were no high IISA achievers. For that analysis, they were all considered to be low achievers. (A more detailed summary that includes a complete list of sample size, bisection point, slope and intercept of all the data sets separately is in Appendix 16.)

7.4.3.1 The privileging effect

This section explains why the student IISA marks rank was chosen as the x-axis of the regression line charts, rather than the GSA marks rank. Where two sets of data points are on the same scatter plot and where the x-axis is the rank order of one of them, the other data trend-line will always be flatter. Privileging one set of data over the other will lead to the privileged data producing the steeper regression line. For example, Figure 29 shows the distribution of the same A1 study data set, including the data points, along different x-axes.

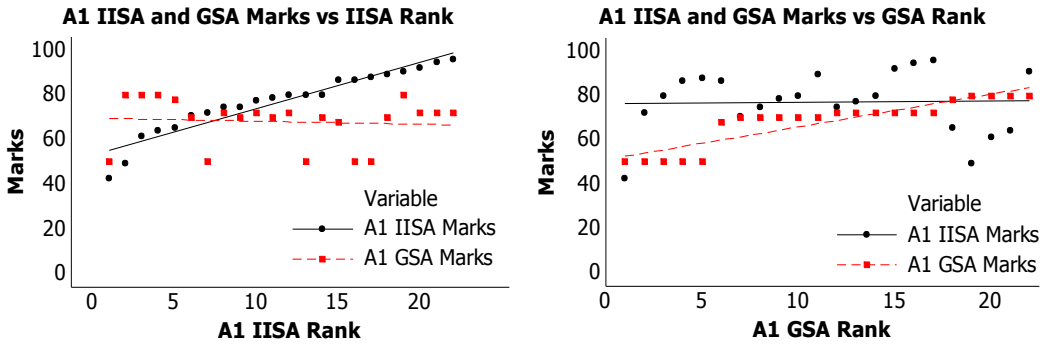


Figure 29: Privileging effect example

For this example, the study data set chosen was, for clarity, the one with the smallest number of data subjects (22).

The focus of this study was on the quantitative impact of GSA mark on students overall marks. It was not about the impact of individual written examination, or of IISA coursework project viva or essay marks. Each IISA and GSA data pair is from the same student. The flatter second regression line is a consequence of the characteristics of the data. It is not simply an outcome from the method of analysis. The relationship between the two regression lines are study findings.

7.4.4 Regression results

7.4.4.1 Marks differences between low and high individual achievers

At the point of the dual-line IISA and GSA regression line bisections, where A is the slope, B is the bisection point and *x* is the distance along the x-axis,

$A_{IISA}x + B_{IISA} = A_{GSA}x + B_{GSA}$, or

$x = (B_{GSA} - B_{IISA}) / (A_{IISA} - A_{GSA})$

For example, the bisection point of a single data set, A1, was at 7.43 along the x-axis:

x = (69.7 - 53.3) / (2.07 - (-) 0.137) = 7.43

This means this study predicted that the first seven lowest IISA ranked students would be advantaged by GSA marking. (See for example, the chart in the left-hand panel of Figure 29. NB: Data set A1 is atypical of the majority of the data in this study.) The reverse was true for students 8 to 22.

For the combined study data, the x value at the bisection point was 3192. This bisection relationship meant that for this combined data, the lowest IISA attaining 78% of students benefitted from GSA marking. On the other hand, it also meant that the top 22% of the individually attaining students were disadvantaged by the marks from the GSA tasks.

Overall, the average of the data sets' GSA marks is higher than the average IISA mark for 82.5% of students at the lower IISA achievement level. Conversely, on average, the highest 17.5% of IISA achievers scored *lower* marks in their group-assessed items than in their individually assessed items, (see Appendix 16).

For the combined data (n = 4069,) the difference between the *average* low IISA mark and the *average* low GSA mark is 6.75. A summary of the differences is shown in Table 24.

Table 21: Low IISA achievers average IISA and GSA marks (Combined data)

Low IISA	Individual Mark	GSA Mark	Difference
mean	56.1	62.9	6.8
SD	9.0	10.06	1.01
count	3192		
max	67.7	100	32.3
min	0	0	0
Range	67.7	100	32.3

For the high IISA students data, (n = 887), the difference between the *average* IISA mark and the *average* GSA mark is -2.2. By this analysis, GSA was shown to disadvantage 887 students in this study. A summary of the differences is listed in Table 25.

Table 22: High IISA achievers average IISA and GSA marks (Combined data)

High IISA	Individual Mark	GSA Mark	Difference
mean	74.1	71.8	-2.2
SD	5.5	8.7	3.2
count	877		
max	96	100	4.0
min	67.7	42	-25.7
Range	28.3	58	29.7

The model showed that all of the students in data sets D11 and D17, where the regression lines bisected each other beyond the range of the number of students in the cohort, benefited from GSA marking. In the remaining 16 data sets, overall the IISA and GSA regression lines bisect 71.3% along the x- axis.

The percentage of the lowest attaining students who benefited by GSA in the separate data sets varied between 34% (A1) and 98 % (D16). These values are 34% and 100% when data sets D11 and D17 are included. The percentage of higher IISA achieving students who were put at a *disadvantage* by GSA is the converse of these figures. They vary between 67% and either 2% for the reduced number of data sets, or none if all the data sets are included.

7.4.4.2 IISA and GSA marks differences between faculties

Table 26 presents the IISA and GSA dual-line charts bisection points of the D data set by faculty.

Table 23: D data IISA and GSA dual-line charts bisection points by faculty

Data Sets	Faculty	Sample Size n	Average Bisection Point as % of IISA Marks Range
D11 and 12	Arts and Humanities	56	166
D13, 14 and 15	Science	1991	73
D16, 17 and 18	Social Science	1574	98

The table shows the bisection point of the GSA and IISA regression lines as a percentage of the IISA range. Group task marking gave an advantage to this percentage of lower individually attaining students in the cohorts. The dual-line charts D16, D17 and D18 in Appendix 12 are of data from the Social Science Faculty. The average of the bisection points of these IISA and GSA regression lines was very close to the top of the IISA attainment range (98%). Data sets D11 and D12, from the Arts and Humanities Faculty are not readily comparable with the other two faculties. The sample size was very small. Figure 30 shows the bisection points of a dual-line chart of the regression lines from both the Science and Social Science Faculty Data.

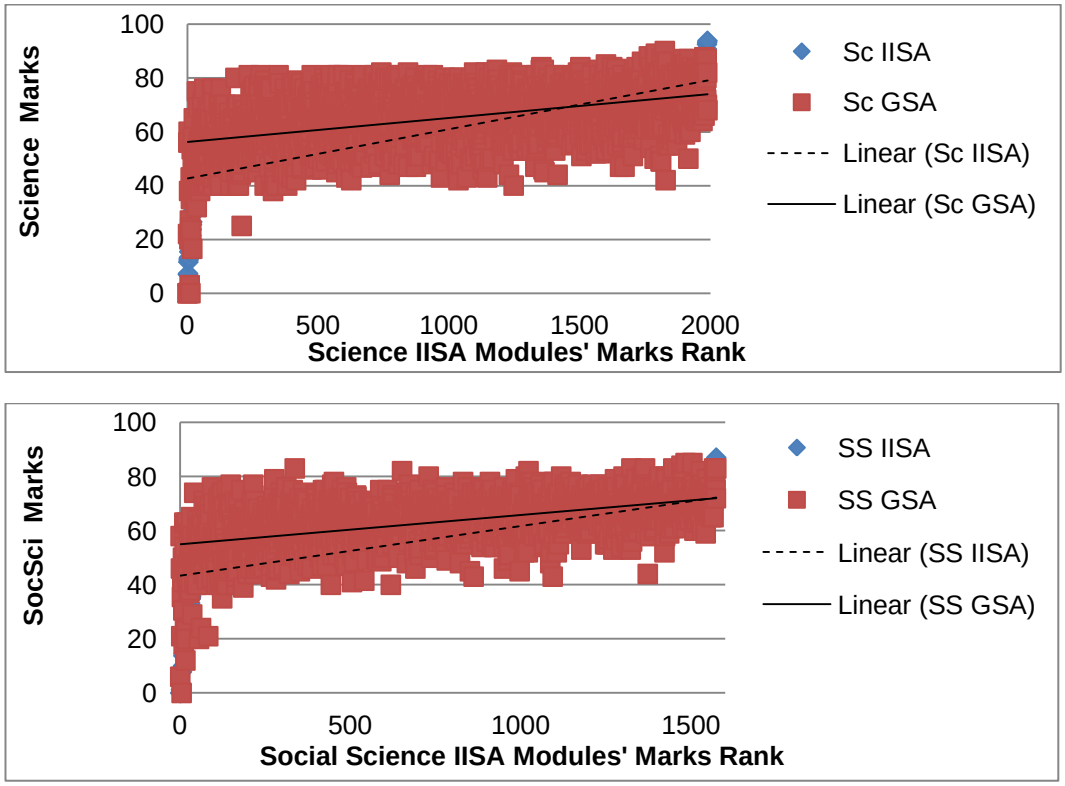


Figure 30: Science and Social Science faculty data regression lines

The charts show that there are differences between distributions of the IISA and GSA categories of marks. In the dual-line charts, the slopes of the IISA regression lines of the two faculties were the same to the fourth decimal place, (0.01841 cf. 0.01839). The slope of the Social Science Faculty GSA regression line was 0.011. The Science Faculty GSA regression line slope was 0.009.

As shown in Table 26 as well as Figure 30, the bisection points are quite different; at 73% of the IISA module marks range for the Science Faculty marks compared to 98% for the Social Science marks.

This finding also supported the unverified finding attributed to Exley and Dennick in section 3.8 (and Wollongong (2006)). In this present study, high achieving Social Science students marks were almost unaffected by GSA marking. The regression lines bisected at 98% of the marks range (for example see Table23). This meant that most students were categorized as lower achieving, and were advantaged by their GSA. (Also, see charts D16, D17 and D18 in Appendix 12).

The distribution of the data points in the dual line regression charts suggests a systematic difference between GSA and IISA marks across the sample. This is explored later in section 7.5, using multilevel modelling which deals with the clustering of students within the data sets and the

differing slopes.

7.5 Multilevel modelling

Multilevel modelling assumes that each dataset is a random sample from a population, and as mentioned in section 5.6, it addresses the shortcomings of simple regression by generating appropriate error terms which are misestimated in simple regression when the clustering is ignored. (Also see section 5.2). It was also used to look for additional evidence supporting the alternative hypothesis H_1, (see Table1). That is, to support the view that there is a systematic difference between GSA and IISA marks in the population and, as with meta-analysis, to look for consistency across the studies. The software used was MLwiN version 2.15. The software option Equations (see example in Figure 31 shown later) was used to model the relationship between the GSA mark and the IISA mark. In this analysis, the dependent variable was the GSA mark and the only independent variable was the individual or IISA mark.

7.5.1 Raw scores multilevel model

This section used MLM to model the raw scores. Results are shown in Table 27.

Table 24: Parameter estimates table: Raw scores data

	Null Model	Model 1 centred round grand mean
Fixed:		
cons (intercept) Slope	66.21(1.75)	66.17(1.44) 0.38(0.05)
Random:		
Student (σ^2_{e0}) Group Intercept (σ^2_{u0}) Group Slope (σ^2_{u1}) Group Covariance (σ_{u01}) Group Correlation r	95.56(2.12) 53.20(18.25)	65.31(1.45) 35.44(12.36) 0.039(0.016) -0.45(0.34) -0.38
2*loglikelihood(IGLS Deviance)		
Fixed	30173	
Random		28660
The IGLS (Iterative Generalised Least Squares) estimation procedure is equivalent to maximum likelihood under Normality. (MLwiN Help v 2.02.03)		

In the table, where the standard error is included with the coefficient, a rule of thumb is that a coefficient is statistically significant at the 95% level if, when divided by the SE, the outcome is equal to or greater than 2 or equal to or less than -2 (Rasbash et al. 2009). For example, the value of the cons (intercept) of the Fixed Null Model is 66.21 and the value of the standard error (SE) is 1.75. The coefficient divided by the SE was 66.21/1.75 = 37.8. The intercept was statistically significantly different from zero because the result was far in excess of ±2.

The Raw Scores Fixed Null Model is the estimate of the average GSA mark without the intervening variable, the student individual mark. The estimated MLM GSA mark outcome was 66.21 with a standard error of 1.75.

The estimated variance of the student marks, (σ^2_{e0}), was 95.56(2.12), statistically highly significant and showing that the students marks were different from one another. The estimated variance of the Group Intercept, (σ^2_{u0}), was also statistically significant at 53.20(18.25) showing that the average scores of the groups were different. It is clear from both the individual students and the group variances that more than a third of the total variance, (53 / (95+53) ≡ 36%) is associated with the group membership but it also varied considerably between individual students.

When compared between models, the -2loglikelihood(IGLS Deviance) gives an indication of the level of improvement, or otherwise, in the model fit. The lower the value, the better the model fit. From Table 27, model 1 which included a predictor variable, was a better fit than the Null model because the -2loglikelihood(IGLS Deviance) was lower, at 28660 compared to 30173 for the fixed null model. The MLwiN Equations window for model 1, the Random Effects Raw Scores model, is shown in Figure 31.

$$\text{GroupMark}_{ij} \sim \text{N}(XB,\ \Omega)$$
$$\text{GroupMark}_{ij} = \beta_{0ij}\text{cons} + \beta_{1j}(\text{IndividualMark-gm})_{ij}$$
$$\beta_{0ij} = 66.171(1.439) + u_{0j} + e_{0ij}$$
$$\beta_{1j} = 0.385(0.052) + u_{1j}$$

$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim \text{N}(0,\ \Omega_u)\ :\ \Omega_u = \begin{bmatrix} 35.439(12.364) & \\ -0.454(0.338) & 0.039(0.016) \end{bmatrix}$$

$$\begin{bmatrix} e_{0ij} \end{bmatrix} \sim \text{N}(0,\ \Omega_e)\ :\ \Omega_e = \begin{bmatrix} 65.307(1.454) \end{bmatrix}$$

*-2*loglikelihood(IGLS Deviance)* = 28659.585(4069 of 4070 cases in use)

Figure 31: Model 1 random effects raw scores MLwiN Equations Window centred round the grand mean

In model 1, centred on the grand mean, the individual mark has been added as the predictor, with the slopes being allowed to vary. The predicted intercept of model 1 is 66.17, making it statistically significant and very close to that of the null model. The predicted intercept standard error is 1.44. The mean slope is 0.38, and the standard error is 0.05, again this is statistically very significant. The variance of the predicted student mean was estimated at 65.31 with a standard error of 1.45, which means that again this is statistically very significant. This compares to 95.56(2.12) for the

null model. The variance of the group, i.e. the level two, module intercept is now reduced to 35.44 from 53 without the predictors, and the standard error is 12.36. This is statistically significant and means that at level two, the data set level, this is unlikely to be a chance result and the data sets are different to one another. This is a reduction of ((53-35)/53) * 100%, or 34%. The variance of the slope is 0.039 and the standard error is 0.016, which is also statistically significant, so the difference between the slopes is unlikely to be by chance.

The covariance of the group, level 2, slope and the intercept is -0.45 with a standard error of 0.34. This is not statistically significant, despite a relationship shown in Figure 31. This also corresponds to a correlation of -0.38. The sign is important. It indicates that as the mean value of the module increases, the slope decreases. As mentioned elsewhere, this flattening, or ceiling effect, can also be seen in the skewed data of set A3 in Figure 14. From this figure, the student marks for data set A3 could not have extended very much higher. They are constrained by the maximum score of 100. The ceiling effect is shown in Figure 32.

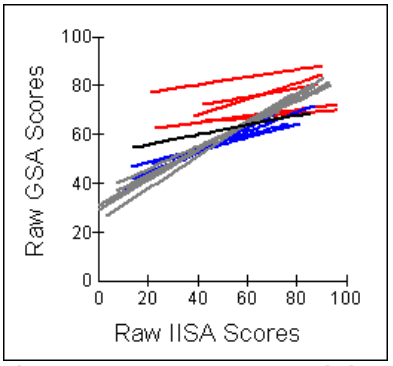


Figure 32: Raw score module predictions

The chart shows the raw data MLwiN generated Residuals charts. It shows the module slopes and intercepts residuals of the 18 raw data sets from Model 1. As shown earlier however, this raw scores relationship is not significant at the 5% level. The difference between the raw data sets slopes and residuals is shown in Figure 33.

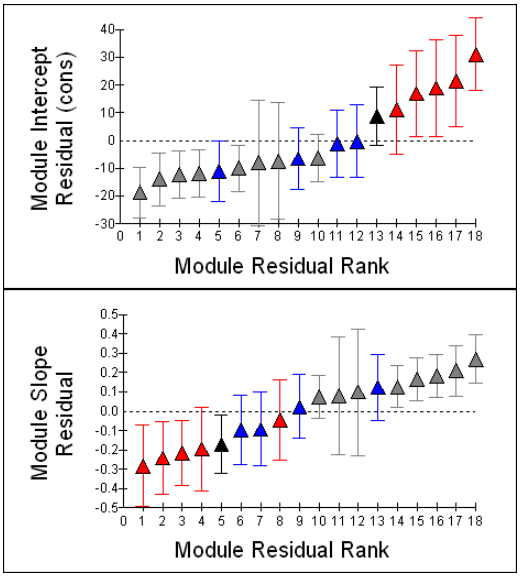


Figure 33: Raw data slopes and intercepts residuals

The rank of the slopes and intercepts, and the sample sizes are shown in Table 28.

Table 25: Raw data residuals rank

Data Set	Raw Data Slopes Rank	Raw Data Intercept (cons) Rank	Sample size	Data Set	Raw Data Slopes Rank	Raw Data Intercept (cons) Rank	Sample size
A1	1	16	22	D11	12	7	25
A2	2	15	26	D12	11	8	31
A3	3	18	48	D13	10	10	581
A4	4	17	49	D14	16	3	658
A5	8	14	74	D15	14	6	750
B6	5	13	53	D16	17	2	285
C7	7	12	35	D17	18	1	445
C8	13	5	53	D18	15	4	844
C9	9	9	38	mean			226
C10	6	11	52				

The slope ranking calculated by the MLM software is different to the ranking from the individual data sets regression analysis (section 7.4). This is because the procedure has 'shrunk' the slopes closer to the overall average slope of the whole data set. This shrinkage is proportional to the reliability of the estimates so it is strongly influenced by the low value of n in the group.

There are broad similarities between the distribution pattern of the data in the Residuals chart, and those of the Forest Plots in Figure 37 (shown later). This multilevel model also found no statistical significance between several of the data sets and the average, including D11 for both intercepts and slopes, as shown in the previous figure. This is indicated by the range of the SE including the zero point. The statistically significant ($p \leq .05$) data in the MLM Residuals charts are, like the Forest Plots, those where the 95% CI does not overlap the zero line.

In this study, the data sets where the intercept and slope are most significantly different from the average are those ranked 16th, 17th and 18th. These are data sets D14 (n = 658), D16 (n = 285) and D17 (n = 445). This is shown by their error bars in the Residuals chart. They are furthest away from the zero point on the y-axis. Additionally, their CIs are short in relation to most of the other data.

On the other hand, some of the data samples were not statistically significantly different from zero, for example data ranked 11th and 12th, i.e. data sets D12 (n = 25) and D11 (n = 31). Their very low number of subject cases generally, not just compared to the rest of the D data set may be a contributing factor to this. The distribution of the module predictions in this MLM may be, at least in part, an artefact of the constricted and skewed range of marks rather than an attribute of the data. That is, the range of marks may be constrained. The minimum mark is zero and the maximum mark is 100. In other words, there is a ceiling on the score (e.g. see section 7.2.1). This suggests that it may be more appropriate to use normalised data in the multilevel model to try to reduce this ceiling effect. The next section describes the normalised data MLM.

7.5.2 Normalised data multilevel model

In the previous section raw score data from the study was used in the MLM. In this section, the data were normalised, i.e. statistically manipulated into a normal distribution with a mean of zero and an SD of one. The data predictions lines of such a model will tend to be clustered on zero, see Figure 35. Through the clustering effect, the ceiling effect of the A data may have been reduced (or even eliminated). The software used to generate the modified data was the MLwiN NSCOres (Normal scores) function. (Also, see Appendix 17 Comparison of MLM Raw and Normalised Data.) The variance of the intercept is calculated at the zero point on the x-axis. Where the raw data level two intercept/slope covariance is negative, the variance of the normalised data will be smaller than for the raw score because the zero point is no longer the left-hand end of the x-axis. The analysis was similar to the previous raw data model. The results of the models are shown in Table 26.

Table 26: Parameter estimates table: Normalised data

	Null Model	Model 1
Fixed:		
cons (intercept) Slope	0.24(0.16)	0.25(0.12) 0.46(0.051)
Random:		
Student (σ^2_{e0})	0.88(0.02)	0.64(0.014)
Group Intercept (σ^2_{u0})	0.47(0.16)	0.26(0.09)
Group Slope (σ^2_{u1})		0.037(0.015)
Group Covariance (σ_{u01})		0.087(0.034)
Group Correlation r	-0.90	
2*loglikelihood(IGLS Deviance)		
Fixed	11103	
Random		9834
As with Table 27, the IGLS (Iterative Generalized Least Squares) estimation procedure is equivalent to maximum likelihood under Normality. (MLwiN Help v 2.02.03)		

The estimated Normalised Data Fixed Null Model MLM GSA intercept outcome is 0.24 with a standard error of 0.16 showing that the data was not statistically significant from zero because the SE is more than half the value of the intercept. The Normalised Data Random Null Model estimated the variance of the student marks as 0.88(0.02) indicating that this is statistically very significant showing that the normalised students marks are different from one another. The estimated variance of the Group Intercept is 0.47(0.16), again indicating that the group intercept is statistically significant showing that the normalised average scores of the groups were different.

The -2*loglikelihood(IGLS Deviance) is 11103. Of the two models, from Table 29, Model 1 was the best fit because the -2*loglikelihood(IGLS Deviance) was the lower, at 9834. The MLwiN Equations window for Model 1, the Normalised Data Random Effects model, is in Figure 34.

$$\text{NormGroupMark}_{Student,\ Module} \sim \text{N}(XB,\ \Omega)$$

$$\text{NormGroupMark}_{Student,\ Module} = \beta_{0Student,\ Module}\text{cons} + \beta_{1Module}\text{NormIndivMark}_{Student,\ Module}$$

$$\beta_{0Student,\ Module} = 0.254(0.123) + u_{0Module} + e_{0Student,\ Module}$$

$$\beta_{1Module} = 0.455(0.051) + u_{1Module}$$

$$\begin{bmatrix} u_{0Module} \\ u_{1Module} \end{bmatrix} \sim \text{N}(0,\ \Omega_u) \ :\ \Omega_u = \begin{bmatrix} 0.256(0.091) & \\ -0.087(0.034) & 0.037(0.015) \end{bmatrix}$$

$$\begin{bmatrix} e_{0Student,\ Module} \end{bmatrix} \sim \text{N}(0,\ \Omega_e) \ :\ \Omega_e = \begin{bmatrix} 0.644(0.014) \end{bmatrix}$$

$$-2*loglikelihood(IGLS\ Deviance) = 9834.047(4069 \text{ of } 4070 \text{ cases in use})$$

Figure 34: Model 1 normalised scores marks data MLwiN Equations window

The predicted intercept of Model 1 is 0.25 with a standard error of 0.12. The predicted slope is 0.46(0.05). Allowing this level two MLM module, the data set, to vary allows the predicted student mean to be estimated at 0.64(0.014). The variance of the group, i.e. the level two module, intercept is 0.26(0.091) and the variance of the slope is 0.037(0.015). The covariance of the slope and the intercept is -0.087(0.034). They are all statistically significant values. This indicates that, unlike the raw data, there is a relationship between the slope and the intercept of the normalised data.

There is still a strong negative Pearson's correlation between the level two intercept and slope (-0.90, see Table 29). This suggests that as the mean value of the module increases, the slope decreases.

The difference between the normalised data sets is shown in the Normalised Data Module predictions charts in Figure 35.

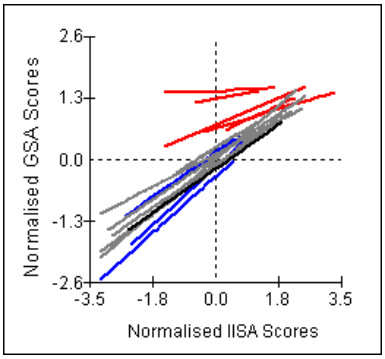


Figure 35: Normalised data module predictions

This shows the slope decreasing as the GSA score increase and suggests a systematic difference in the relationship between the GSA and the IISA marks in the population, which unlike the raw score data is statistically significant at the 95% level. It disproves hypothesis H_0, that the two summative assessment methods have the same impact on students' overall marks. It does not disprove alternative hypothesis H_1, that they each have a different impact on students' overall marks. Figure 36 is the MLwiN Residuals chart of the slopes for the 18 normalised data sets.

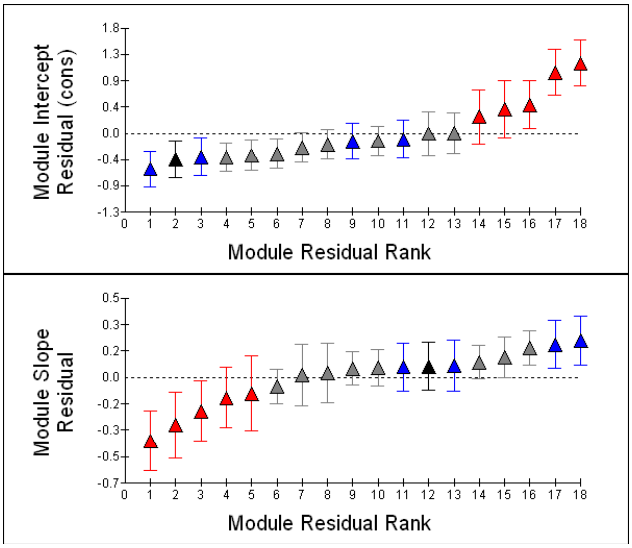


Figure 36: Normalised data slope residuals

The rank of the slopes and intercepts, and the sample sizes are in Table 27.

Table 27: Normalised data slope residuals rank and data set ID

Data Set	Normalised Data Slopes Rank	Normalised Data Intercept (cons) Rank	Data Set	Normalised Data Slopes Rank	Normalised Data Intercept (cons) Rank
A1	1	18	C10	11	9
A2	2	17	D11	7	12
A3	3	15	D12	8	13
A4	5	14	D13	6	10
A5	4	16	D14	16	7
B6	12	2	D15	9	8
C7	13	11	D16	15	5
C8	17	1	D17	10	6
C9	18	3	D18	14	4

For sample size, see Table 28.

By observation, from Figure 32 and Figure 35, normalising the data seems to have had little impact on reducing the ceiling effect. The red A data prediction lines remain flatter than the others and still predict higher average individual and group scores for the students. In a table of the comparison between raw and normalised slope variance values (Appendix 17), the ranking only varies slightly between them. The ranking of the slope variance only varied slightly between raw and normalised data). This model of normalised data also found no statistical significance in several of the data sets, including D11, ranked 7th (see Figure 36).

7.6 Meta-analysis

As noted earlier in section 5.6, meta analysis assumes that each data set is from a separate experiment and combines the data from all of these *'separate experiments'* to estimate their combined robustness, i.e. to check that each "experiment" is producing statistically indistinguishable results. Part of the output of the software used for analysis is forest plots which

give a picture of the results and allow a visual confirmation of the statistics. In this section, the average scores from IISA and GSA are compared across the different data sets.

Meta-analysis (MA) is one of the methods used in quantitative systematic reviews. It allows the results from a sample of studies, each of which might not have had very robust results alone; to be combined to produce robust results overall. It uses the effect size of the individual studies as the common variable. Effect size is the difference between the means of the two groups, divided by the standard deviation. This has been defined as *"the standardised mean difference between the two groups"* (Coe 2002).

Meta-analysis uses Forest Plots to present systematic review data. This section of the thesis uses them to illustrate the differences between the data sets and the overall mean. The first Forest Plot shows the full range of eighteen data sets. Subsequently, a comparison is made of the differences between, and means in, data from both high and low ability students.

Heterogeneity refers to the level of dissimilarity between the samples, e.g. between the group assignment methods or the study teaching and assessment approaches. Each set of marks data for this study will be context specific to the module from which it was collected. In meta-analysis, it is therefore important to explore the heterogeneity of the effect sizes of the sample of studies. It has been noted that:

> *"The test for heterogeneity, Q test, is used to assess the null hypothesis that effect sizes are homogeneous and that sampling error accounts for the variation rather than random variability from other sources in the population of effect sizes. This Q statistic has an approximate chi-square distribution with k -1 degrees of freedom, where k is the number of independent effect sizes."*

(Penny and Coe 2004:225)

Selecting the Meta-analysis option from the Numerical output menu of the MIX 1.7 software produced the meta-analysis results, including the Q statistic and its 2-tailed p-value. In this study, meta-analysis is used to model the mean difference (MD in Figure 37) between the IISA and GSA marks in the separate data sets. The software also calculates the overall difference between the two marks categories. This shows the consistency of the difference between the IISA and GSA marks across the study data sets. Figure 37 shows an annotated Forest plot of the fixed effects model that was generated by the MIX software. This plot includes all the eighteen data sets.

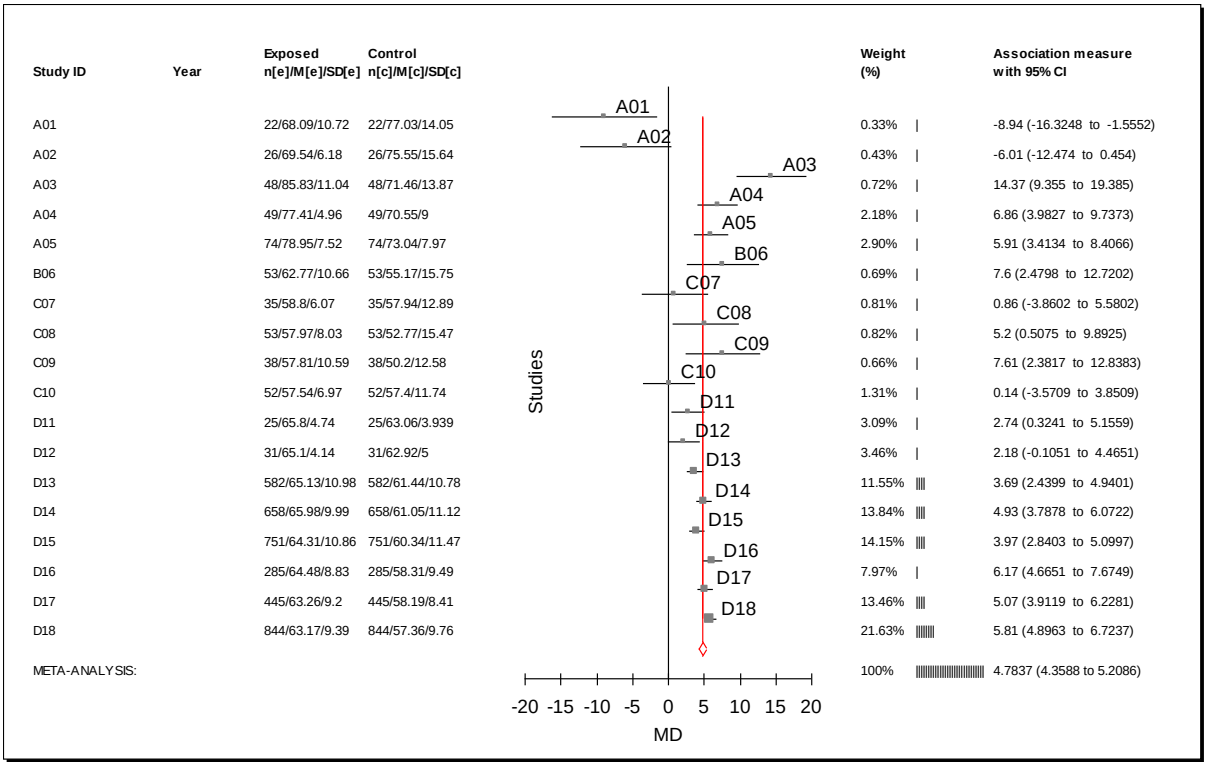


Figure 37: Annotated meta-analysis Forest Plot of the mean marks of all eighteen data sets

For each 'study' or data set, the software application required a study ID and the sample size (n), mean value (M) and standard deviation (SD) of the 'Exposed' (e), i.e. the GSA data, and 'Control' (c), i.e. IISA data to be entered. This is the data on the left of the graphic in Figure 37.

In this software, the study year is an optional data entry item. It serves no purpose in this analysis. The output from the software includes the graphic and the data on the right of it. Although meta-analysis uses the standardised mean difference, i.e. the effect size, as the measure of comparison, this is not entered directly into the software, neither is it presented in the Forrest Plots.

The heterogeneity statistic Q had a value of 73 and a two-tailed p-value of < .005. The meta-analysis MD (mean difference) was 4.8 as shown by the red vertical line with the rhombus at the bottom end of the line. The low p-value from this data is evidence of heterogeneity in the sample. Table 28 lists the differences between the data set MD and the meta-analysis mean.

Table 28: Differences between the data set MD and the meta-analysis mean

Data Set	Forest Plot MD	m - MD	Data Set	Forest Plot MD	m - MD
A1	-8.9	13.7	D11	2.7	2.0
A2	-6.0	10.8	D12	2.2	2.6
A3	14.4	-9.6	D13	3.7	1.1
A4	6.9	-2.1	D14	4.9	-0.1
A5	5.9	-1.1	D15	4.0	0.8
B6	7.6	-2.8	D16	6.2	-1.4
C7	0.9	3.9	D17	5.1	-0.3
C8	5.2	-0.4	D18	5.8	-1.0
C9	7.6	-2.8	mean	4.8	
C10	0.1	4.6			

There are several candidates for contributors to the heterogeneity of the sample. For example, the difference between the A1, A2 and A3 data sets and the average for the sample is 13.7, 10.8 and 9.6 respectively, compared to the next highest difference, data set C10 (4.6). When these data sets are removed from the model, however, the evidence of heterogeneity, the p value, remains almost as strong. The Q statistic had a much lower value of 34.8 although the two-tailed p-value of remained low at 0.002. The meta-analysis MD is still 4.8 to two significant figures. This is shown in Figure 38.

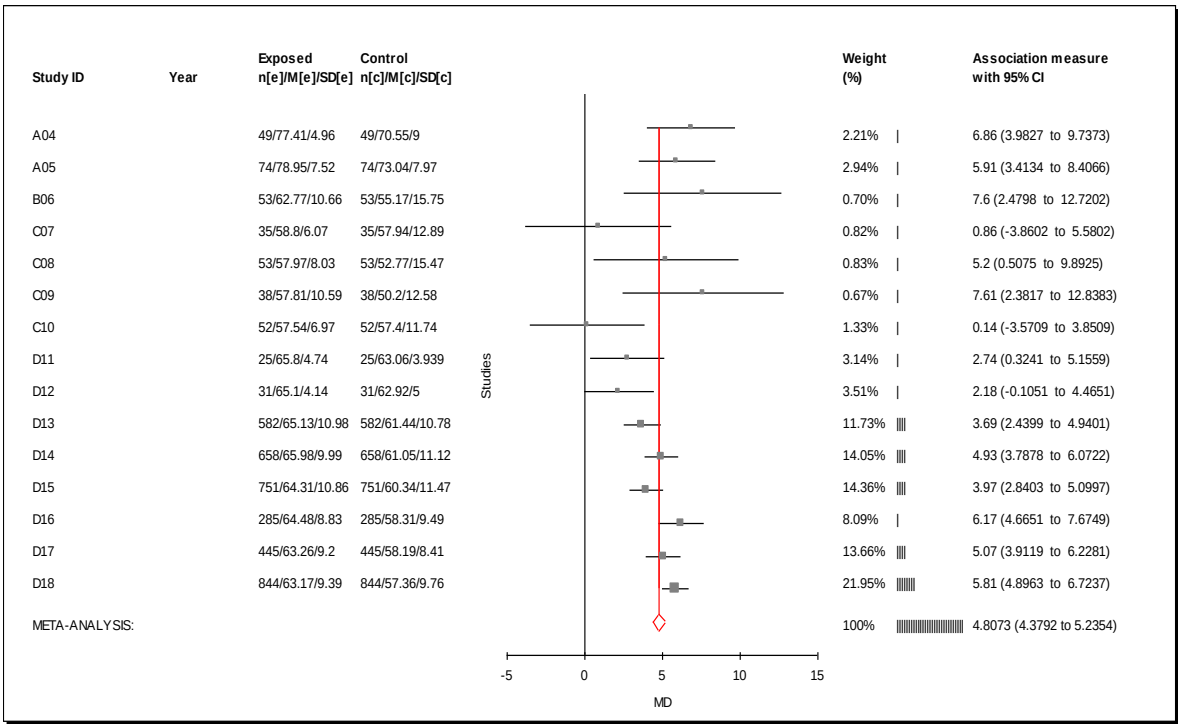


Figure 38: Fifteen study Forest Plot

Removing the A1, A2 and A3 data from the analysis reduced heterogeneity but it remained high. Fletcher (2007) has noted that there are two approaches to dealing with heterogeneity in meta-analysis. The first one was to *"avoid summarising the result and look for reasons for the heterogeneity"*. The second was *"using another method—the random effects model"* (Fletcher 2007). This was used in the previous section, multilevel modelling.

In an attempt to “look for reasons”, analyses were run separately for individually low and high achieving students, i.e. those to the left and right respectively of the intersection on the dual-line regression chart, (see Figure 26). Data sets D11 and D17 were excluded from this analysis because all the students were ‘low’ achievers (see for example section 7.4). The results are shown in Figure 39.

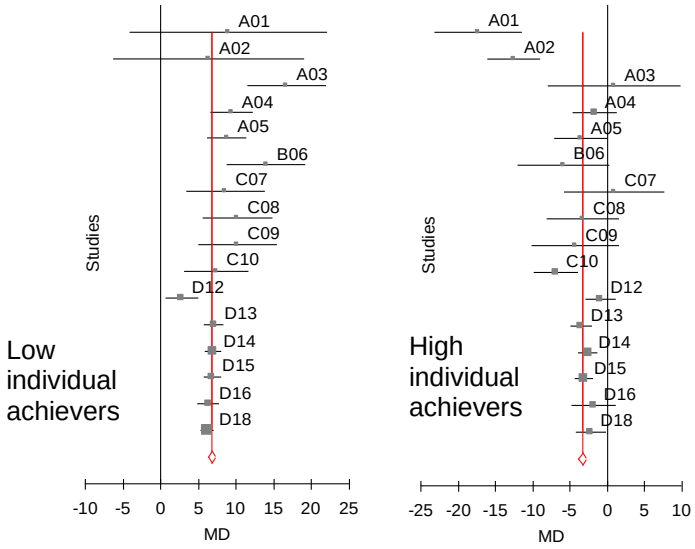


Figure 39: Study data Forest Plots of differences between low and high individual achievers IISA and GSA mean marks

Figure 39 shows two un-annotated (standard) Forest Plots produced by the MIX version 1.7 MA software. The left-hand plot shows the effect that GSA has on the average low IISA achiever. The MD is 6.8. The right-hand plot shows the effect that GSA has on the average high IISA achiever. The MD was minus 3.3. By this analysis, the difference that the two assessment methods have between the average high and average low IISA student was more than10 marks (6.8 - -3.3).

There are 18 data points in Figure 38. There are only 16 in each of the figures in Figure 39. This is because the bisection point of the GSA and IISA regression lines of the dual-line chart for two of the data sets, D11 and D17, (see section 7.4,) occurred beyond the high end of the IISA rank. (See e.g. Appendix 16.) In these two data sets, all students were low achievers according to the regression model. As there would be no high achievers to compare them to, they were omitted from the low and high IISA achievers versions of the Forest plot.

Data set A2 had an average MD of -6.01 with a range from -12.47 to 0.45. Data set C7 on the other hand, had a mean of 0.86 and a range from -3.8 to 5.58. Data set C10 had an MD of 0.14

and a range from -3.57 to 3.85. In this MA model, the software calculated that these data sets were not statistically significant because the CIs include zero.

As described in the opening paragraphs of this section, in the MIX MA software nomenclature, e was the label given to the data from the exposed cohort. For this meta-analysis, this was the marks for the GSA item. The IISA group was labelled c. It was the control group. The difference between them, i.e. M(e) minus M(c), was 4.78. The 95% confidence interval was from 4.36 to 5.21. For a student, this could mean the difference between two adjacent degree classifications.

7.7 Simulation

The aim of this simulation exercise was to analyse uncorrelated simulated data to test the hypothesis that uncorrelated data would produce the same IISA/GSA regression lines crossover pattern seen in correlated data. It did. This hypothesis is alternative hypothesis H_2 from Table 1.

The data was simulated by using the Calc, Random Data, and Normal options from the Minitab statistical software (r14) menu. The number of data cases was selected at 100. The individual mean mark of 50 was chosen with a normal distribution. The selected group mean mark was 60. The selected standard deviation for the individual simulated marks was 10. For the GSA marks, it was five. This approximates to the typical higher mean and lower SD for GSA data compared to IISA marks that was reported by Downie (2001) and found in this study. The two sets of data were unconnected. Figure 40, Figure 41 and Figure 42 are scatterplots and a histogram of this uncorrelated simulation based around those found in data sources C and D.

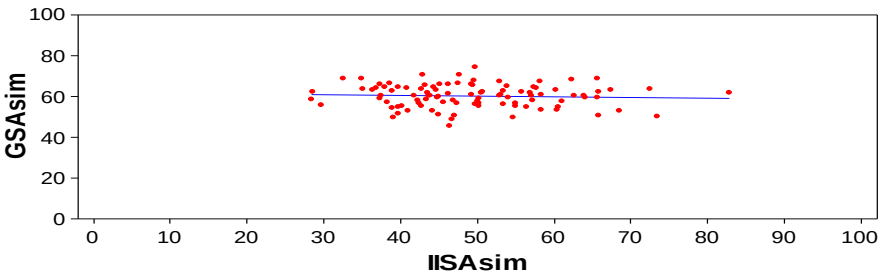


Figure 40: Uncorrelated simulated data single line regression chart

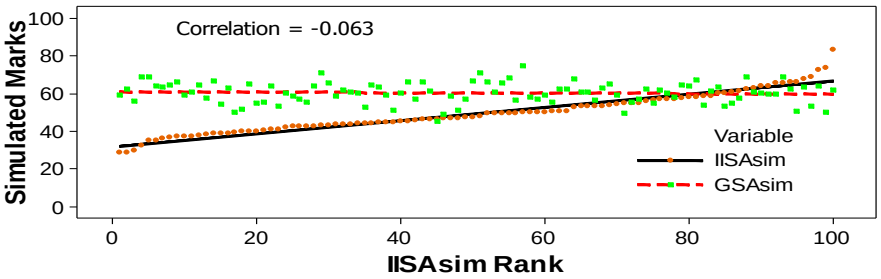


Figure 41: Uncorrelated simulation marks scatterplot with regression lines

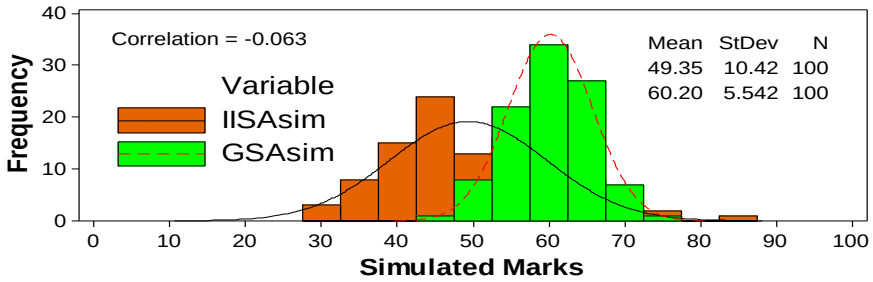


Figure 42: Uncorrelated simulation marks distribution histograms

As the data were simulated, there was no step in the histograms below, and spike at, the 40 mark as was reported for data from source D (Figure 21). Neither was there a small cluster of data at and around zero. The simulation histograms of normally distributed values were also a better fit to the normal curves than the study data. In spite of these differences, the patterns of this synthesized data closely mimic that of the study data.

In this simulation, almost 20% of the highest individual achievers would have been disadvantaged by these GSA marks (see Figure 41). On the other hand, more than 80% of students in this situation would have benefitted by higher scores if the assessment item had been scored by GSA marking.

The alternative hypothesis H_2 (see Table 1) was not disproved.

7.8 Data analysis chapter summary

This chapter has presented the study data analysis and results. This study has shown that, on average, for the low IISA achievers, their GSA marks were higher than their IISA marks. The reverse was also true. In addition, the difference between the IISA and GSA marks of the low achievers was greater than the difference for the high achievers and in the opposite direction. The alternative hypothesis H_1 (see Table 1) was not disproved. The chapter is further summarized below under its main section headings.

7.8.1 Data from four sources summary

See sections 7.2.1 to 7.2.4. A separate data sources summary was also presented in section 7.2.5. This section presented an analysis of the data from the individual data sources.

All of the data distributions approached normality although there were differences between the A data and the data from the other sources. Only three of the A data sets supported the Downie (2001:7) finding that the higher GSA mean value with a lower standard deviation is typical of this type of data. In addition, some of the A data sets exhibited a ceiling effect. The dual-line regression chart from the reanalysed Knight (2004) module marks B6 data set showed the same high IISA achievers lower GSA; low IISA achievers higher GSA scores pattern as that found in the other data in this study.

Data source C was the single Science Faculty module, which was specifically designed to include providing undergraduates with group working experience. Data was collected at the lowest marking level, i.e. the assessment item level. From this, analyse was possible at both assessment item level and at module level. Unsurprisingly the analysis of scores at coursework item level showed a greater correlation between the IISA and GSA marks than between IISA and GSA scores at module level. The D data, which was separated into eight data sets, was around 79% of the data used in this present study. The distribution of the D data showed a very high peak at the 40 mark. Reasons for this were discussed. All the B, C and D data supported the Downie (2001) finding.

7.8.2 Correlation summary

See section 7.3. Alternative hypothesis $H_{3,}$ that there will be a low correlation between IISA and GSA category marks (see Table 1), was explored. The correlation between the IISA and GSA marks is a key relationship in this present study. Correlations were divided into three categories: Low: <= 0.33, Medium: 0.34 to 0.66, and High: 0.67 to 1. All the *module marks* data sets have either a medium or a high correlation.

All correlations between data sets at the *coursework element* level, i.e. data sets A, B and C, were low. At this level of marks data, the correlation was the least corrupted because they did not have IISA item marks added to their total. Due to this dilution effect, although the modules correlations did not support the alternative hypothesis H_3, neither was it disproved. This supports the hypothesis that, overall, IISA and GSA were measuring different constructs or that one or both

measures unreliably.

From source C GSA module, the *coursework elements* Spearman's rho correlations were low and ranged from 0.16 to 0.31. The correlations of the C data *module* marks on the other hand, were from 0.34 to 0.75. This supports the hypothesis that, while the module marks measured similar constructs, overall, IISA and GSA were measuring different constructs or that one or both measures were unreliable.

7.8.3 Regression summary

See section 7.4. These were single- and dual-line models. In the single-line chart (e.g. Figure 24,) the spread and slope of the module item marks allowed comparison between data sets. The dual-line model showed both the IISA and GSA marks, in the y-axis, against the IISA rank on the x-axis, (e.g. see Appendix 12). It allowed comparisons between students within data sets. The models also showed that, although there were some exceptions in the data, overall, GSA marking would result in higher marks than the IISA marking for the lowest IISA achieving students. Conversely, they predicted that GSA would disadvantage students who were the highest IISA achievers. They also showed that the disadvantaging of high achievers was less than the advantaging effect on low achievers and that the effect was faculty specific.

The regression analysis indicated that for example, if each module item had been assessed using GSA, then the marks of the highest IISA attaining student in the A1 data set would have been lower by 32%. In addition, it predicted that the lowest individual achiever in data set C8 would score only 30 marks for the IISA item. They would score 50 for the GSA item. They would benefit from GSA by 64%. Ensuring equality in IISA marking from GSA marking could be problematic.

7.8.4 Multilevel modelling summary

See section 7.5. The Raw Scores Fixed Null Model estimated intercept was 66.21 with a standard error of 1.75. In model 1, centred on the grand mean, with the individual mark added as the predictor and the slopes being allowed to vary, the predicted intercept of model 1 is 66.17. Fixing the level one variable (student) and allowing level two (data set module) to vary allowed the predicted student mean to be estimated at 65.31(145) compared to 95.56(2.12) for the null model. The variance of the group was also reduced, to 35.44, from 53 without the predictors. In the raw scores multilevel Model 1 the covariance of the IISA and GSA prediction was negative. This indicated that the raw data included a ceiling effect, although this was not statistically significant for the raw scores. It resulted in the higher prediction lines being flatter. The attempt to correct this

statistically by normalising the data was presented. The outcome remained largely unchanged, although the result from the normalised data was statistically significant. In addition, the change in the ranking of the data sets between the raw and normalised data Residuals charts was small.

More than a third of the total variance is associated with the group membership but it also varied considerably between individual students. The coefficients of the variances of the slopes showed that there was a substantive difference between the A data sets and the others. The analysis suggested a systematic difference in the relationship between the modules in the population that is statistically significant at the 95% level. Several of the data sets were not statistically significant.

7.8.5 Meta-analysis summary

See section 7.6. Forest plots were presented to illustrate a meta-analysis of the differences between the data. The mean difference between the mean GSA and IISA marks from all eighteen data sets was 4.8. This supported the overall difference (where n, the total number of students in the study = 4069) reported in section 7.2.5. The effect of GSA marking on the average low IISA achiever was 6.8. The effect on the average high IISA achiever was -3.3. This showed that by the meta-analysis, GSA disadvantaged the highest individual achievers by a little less than half the amount that the lowest achievers were advantaged, which confirmed the previous analysis.

7.8.6 Simulation summary

See section 7.7. The simulation analysis supported hypothesis H_2 (Table 1). Both types of simulated data scatterplot, and the marks histograms, had a similar format to those that showed actual data. The scatterplots of actual and simulated data, in section 7.7, had similar means, standard deviations and sample sizes. On the other hand, the distribution of the data points was different. The simulated GSA data points were more scattered than the actual GSA data points, i.e. in the dual-line charts, the residuals of the simulated unprivileged data in the dual-line chart were more varied.

The next chapter presents discussion on the study, the data, and on the study findings. It also includes discussion on other issues of interest that arose during the exploration.

Chapter 8. Discussion

The previous chapter described the present study data processing, analysis and findings.

More usually, the discussion chapter concerns the issues building from the literature review. As mentioned earlier however, there were very few studies found on the topic and these were small studies or analysis as a small part of a different study. The first section of this chapter reflects issues arising from the literature review. It discusses the review of the literature that formed the general education background to this present study. The second section is reflective, based on the study and on background reading that while relevant, did not readily fit elsewhere in this account. It also relates to study limitations, conclusions, and, by recommendations for further study, other issues raised by this study.

8.1 Discussion 1: Study background literature review

This section will discuss issues arising from the general background literature review (chapter 3).

8.1.1 Group or team

As mentioned earlier, in the literature, the distinction between the terms *group* and *team* is often unclear, even interchangeable (e.g. see Much 1998 and section 3.1).

By the Sheard and Kakabadse (2004:13) definition, student GSA groups are teams, but most universities seem to prefer the word *group* to a work unit consisting of two or more people working cooperatively towards a common aim. Whatever they are, the impact that their collective summative assessment marks have on their individual marks is the reason for this present study. S*tudent groupwork teams* may seem to be a somewhat pedantic phrase. It might however help to describe and explain the student team/group name dilemma.

As noted earlier, the preferred term in the world of employment is *team*. The term *team building* also seems to be more common than the term *group building*. There are commercially run courses specifically to promote *team building* in the client organizations. There are no courses purporting to promote *group building* for example. In addition, of course, whoever heard of the term *group-spirit*?

Pragmatically, a group can be a collection of anything. They can be animate or inanimate. We do not refer to a team of bone china figurines, for example. Referring to them as a group, rather than a collection, is acceptable. A team must be two or more animate entities, e.g. a team of oxen

pulling a plough, or a team of people engaged in some mutual aim.

8.1.2 Practise effect

The practise effect (see section 3.3,) is not specific to student GSA projects. It does however add another level of complexity to any analysis of the GSA effect. The more GSA modules a student studies, the more experience of it they will have. This will include the opportunity to learn and practise interpersonal skills. From their second GSA module onwards, this will give those who experience multiple GSA a possible advantage over those who only study one GSA module in their programme. Low IISA achievers who study multiple GSA modules will be at a greater advantage than those who only study one GSA module. Similarly, high individual achievers would have an opportunity to learn to mitigate the disadvantages that GSA holds for them.

From the GSA modules taught at the university in northeast England, it is clear that some schools or departments taught only one GSA module that was eligible for this study (see section 5.1 for example), while others taught more. Students assessed by GSA methods for more than one module would have prior or concurrent experience of the *forming, storming* and *norming* (section 3.9.5). They would therefore have the potential to speed up these activities giving them an advantage by leaving more time for *'performing'*. For example, from this study, Engineering students could have more experience of GSA than students of most other disciplines could. The effect of the additional practise could mean that they had higher GSA marks than, for example, their Archaeology counterparts. The data as received were anonymised with respect to discipline so there could be no analysis of this or of the effect of multiple weighting of modules (see later).

Additional practise may also affect the GSA marks differences between the low and high IISA achievers. An analysis where all else was equal would be difficult because different disciplines will often have different assessment items and methods. They could also be context specific, with project assessments based indoors or outdoors, or laboratory or field, etc. Double or even treble weighted modules would exacerbate the effect of GSA on students overall marks. (Also, see sections section 2.6.1 and sections 5.1.4 and 6.3.4).

8.1.3 Rationale for practice issues

If we only assess the products of the group efforts in GSA modules, and do not assess students generic skills, this would send a message that learning generic skills was not an important part of the degree programme. If GSA is practiced, whether to teach generic group working skills or for other reasons, then summatively assessing these skills is also important. To ensure fairness to

each student, one would need to ensure that each student worked with a sufficiently large representative sample of his or her peers in different groups. The number could be determined using existing statistical theory. A key point in HEI GSA however, is that the potential partners are other students and they may not reflect the persons that they might meet in the workplace. For example, in the workplace the group may be made up from persons from different disciplines, as well as widely differing aptitudes and attitudes.

Student group work is self-evidently dynamic. It is also a high-stakes and time-constrained environment. It seems likely that some groups could spend too long in explaining, discussing and negotiating their point of view to each of the other group members. Because of this, they may only complete a superficial deliverable product.

8.2 Discussion 2 – on reflection

This final section in chapter 8 reflects on more general education issues, which have a less easily defined impact on this present study, or are not an easy fit into other sections of this thesis.

8.2.1 The effect of teaching and learning on students

For the most part, however a programme module is delivered, whether by lecture, seminar, tutorial, laboratory, fieldwork or any other method, the module teaching sessions are all made available the same way to each student. They cannot individually negotiate the method by which they will be taught. (This does not include special access needs). On the other hand, how it affects students will vary will depend on the effectiveness of students coping strategies, their personality, circumstances and overall ability. Any attempt to deliver a module, or a module item, by a novel or more efficient method, e.g. GSA, will apply equally to the whole cohort irrespective of individual ability.

8.2.2 Data source A

Attempts were made to contact the academic who made the A data available, to query the differences found between the Australian data and the other data sets. The reply was only an automated response. The text indicated that they had since taken on an adjunct role at their university. They had retired from full-time teaching.

8.2.3 High-stakes, consequences and fairness

Any summative assessment could be high-stakes for the assessee and their stakeholders. It may also be competitive. This is irrespective of the level of the assessment. For example, it could be for a pre-school child being assessed for a place in a particularly prestigious kindergarten, or it may

include students taking GCSEs and their carers. They may be concerned for career or further education opportunities being dependent on the assessment results. It may even apply to a candidate for Fellowship of the Royal Society.

An undergraduate's first-degree is however, a high-stakes qualification for several reasons. First, the label university degree is a familiar one. It is well known and likely to be respected by most of society. Second, its result persists throughout the graduate's lifetime, in particular, throughout their working life. Third, a degree classification is often applied as a first filter in employment recruitment selection. It also serves the same purpose in selection for entry to advanced study opportunities.

The consequences of incorrectly assessing student achievement through the misuse of GSA could affect everyone. A highly able graduate could receive a lesser classification. They may miss opportunities in training, employment and promotion. They may even be declined employment at the level appropriate to their abilities.

The GSA method could improperly benefit students of modest ability. They may gain a licence to practice through their marks being elevated because of it. Even if this was only practice at an entry level, the results could be catastrophic. Consider for example wrongly licensed members of the medical, accountancy and legal professions, or wrongly licensed airline pilots and air traffic controllers. The effect of awarding inaccurate degree classifications can and will have far-reaching consequences. Before these conclusions could be supported, the study of the impact of GSA would need to be extended to include its impact on degree classification awards. For this present study, the data on degree classification awards was incomplete.

8.2.4 Peer teaching and assessment

The issue of peer teaching and assessment focuses on the different standards of teaching that seem likely between students and academics.

Some GSA data in the study included results from marking that included peer-assessment. This occurred particularly in source C data; although it was not 100% peer assessed. Peer-assessment is unidentifiable in the remainder of the data.

There were two main concerns with peer-assessment marking. The first was about the assessment of the assessment skills of the peer assessors. The second was peer marking of effort

rather than attainment (see the Taylor and Rumpus (1997) quote in section 3.10). The problem was; should summative assessment be marked on the quantity or the quality of that effort, or both?

Students study at universities for a variety of reasons. For the majority of them, these will include obtaining a degree at the end of their studies, (see Appendix 2). Before beginning their studies, they may not necessarily expect that it will involve peer- and/or self-assessment. Informal consultation of prospectuses and websites suggested that this level of detail is not usually available to prospective undergraduates. In addition, students may have neither the time nor the inclination to learn to be peer assessors. That task arguably, could take a lifetime to master, even if it might only take one seminar session to learn.

8.2.5 Assessing the assessors

The problem with peer-assessment (and with self-assessment) is one of validity. When either of these is included in any module, not just GSA, and there is no evidence that they can do it already, then students are unlikely to learn how to do it unless they have an adequate opportunity to practice it. They would then need to be summatively assessed in it, as part of their degree programme. From the literature review, in section 4.5.6, it seems that they are unlikely to take this part of their programme seriously unless it *is* summatively assessed. There would also be a side issue here regarding the treatment of candidates who do not pass their own peer-assessment assessment. Not summatively assessing their peer-assessment skills sends the wrong message to students. This message is that while peer-assessment is sound enough to practice, it is not important enough for those practicing it to be summatively assessed in their attainment in it.

It is likely that even teachers new to higher education will very quickly become accustomed to assessing students work. They would acquire this from immediate and regular practical experience. On the other hand, student exposure to self- and peer-assessment seems likely to remain sporadic and limited. It would depend on the number of modules in their programme that used peer-assessment.

Academic markers also mark and compare the whole cohort rather than just assessing themselves or other members of their, usually quite small, team. Additionally, academic assessors have their own assessment skills assessed by having a sample of their assessments ‘double marked’ by a peer, but student assessors do not (see section 4.5.6).

8.2.6 Should peer-assessed marks be awarded for effort, quality, or both

There is a serious weakness in some of the methods advocated in the literature on methods of deriving individual marks in GSA modules, (also see section 4.5.5). Peer marks seem only to be awarded for effort (e.g. see section 4.5.2 and Burd et al. (2003), Gibbs (2009) and Taylor and Rumpus (1997)). They do not generally seem to be awarded for the combined quantity and quality of the student's contribution. The outcome or output quality and quantity are both functions of effort and ability. An algorithm based on both to calculate a mark for the *outcome* of that effort and ability is crucial. In the effort only scenario, a student could contribute 100% effort in completely the wrong direction. They could actually contribute nothing to the group assessment product.

Gibbs seems to have been the only researcher to have addressed this issue in print, (see for example section 4.5.2 and (Gibbs 2009)). In addition, his term *'different kinds of contribution'* is also crucial to group assessment. For example, how are the contributions from different group roles, e.g. technical, creative and administrative to be compared, and marked?

By only addressing methods for deriving an individual module mark from the group score and the students contribution of effort, (also, see section 4.5.5,) reports in the literature do not go far enough. While their texts might imply *quality* of effort, there was no publication found with *quality*, rather than *quantity*, specified directly as the important aspect of effort. Nicholl and Alexander (2004a) for example, only emphasise effort. They had nothing to say of the products of that effort. They even provide a marking scale to peer-assess this *effort*. Also, see Conway et al. (1993), and Kaufman, and Felder (2000) for additional examples of this emphasis on effort alone.

This raises several questions. First, what does effort mean anyway? Second, is it the resource or energy expended during the completion of a project with the opportunity cost of not expending those resources on some other project? Third, does the word effort somehow mean something different to the everyday usage when applied to GSA peer assessment? Fourth, what are the implications of this for IISA modules, should they also be assessed by effort alone? If this were to happen, then almost anyone could be awarded a first-degree class classification. They would not need ability, only tenacity.

The priority of the group should be a successful outcome of the group project, no matter how the concept is understood and operationalized. From this point of view, it does not matter if a group

project succeeds through minimal effort from one individual, from maximum effort from another individual, or from maximum effort from all the members of the team. Equally, a group member may be exceptionally enthusiastic, able, articulate and vocal. The others may not share their enthusiasm. They may marginalize the more vocal team member, or even ignore them completely. Alternatively, the more voluble group member may also get their way for the sake of project progress and group harmony. Both of these outcomes would seem to be equally wrong educationally.

This section has focussed on peer assessment. How university teachers practice GSA was not part of it. The need to assess the peer assessors, together with the effort only issues need academic recognition and urgent attention.

8.2.7 Fairness in allocating individual marks in GSA modules

There was no universally accepted method of deriving individual marks fairly from GSA items found in the educational research literature. Neither was anything found for a similarly acceptable principle of fairness in GSA. This is despite the plethora of marking algorithms put forward by, for example, the authors in Appendix 5. The lack of consensus may be the result of tensions caused by the conflicting and, possibly at times, mutually exclusive aims of using groupwork in HE modules. Collaborative working and collective assessment may also conflict with students' individual competitiveness. The goal of finding a universally accepted method of deriving individual marks fairly from GSA items may simply be unattainable. (Also, see sections 2.7.3 and 4.9).

IISA is an established, and in general, a universally accepted, or at least recognised, category of assessment methods for marking independent individual student attainment. It is adaptable to suit the local conditions. Any differences in methods seem likely to be small in comparison to the main difference between GSA and IISA marking categories. When a similarly acceptable GSA category is defined, this will help to legitimise the assessment technique. It will allow those who study under its regimen to understand it more thoroughly, as they might now do with IISA.

8.2.8 Evidence based practice

Consider a pharmaceutical company wanting to market a new drug combination. They publicly assert it will be greatly beneficial to a large section of humanity because it is far more efficient than existing options. Its use has additional benefits for the user. It is also less costly to deliver. Its effectiveness is however unresearched, so development costs were minimal. The company assumes that the new product is the same as existing products, with the same side effects. They

also firmly believe that selling this product instead of the range of well-known products that fulfil similar functions will result in significant cost savings for the company. Would anyone buy such a drug? Who knows? Fortunately, in the case of pharmaceuticals there are legal safeguards, which would forbid such a move.

How is this very different from what has happened from the widespread popularity of GSA (i.e. the treatment) in higher education? The quantitative impact of GSA on an individual's overall marks, and therefore on their degree classification is greatly under-reported. It seems almost as though those who advocate the use of GSA in HE expect stakeholders to buy into this regimen unquestioningly, without having fully researched its effects, much less into its efficacy.

Additionally, in the United States the No Child Left Behind act forbids the use of federal funds on educational initiatives, which do not have scientific backing. The law is meant to move education in the same way that medicine was moved by forcing it to become evidence based. Perhaps this concept should be universally adopted.

8.2.9 Withholding key findings from group peers

Students could gain an advantage over the rest of their group by keeping what they consider key findings from their own project research exclusively for their own use. Almost nothing has been found in the literature that indicated that group members were deliberately doing this. The exception was the Myers-Briggs theory on what extraverts think about introverts. (This was reported in section 3.9.7.1.) It seems unlikely that this situation does not occur, at least occasionally. The stakes are very high.

Some students may think that they could improve their marks, and subsequently their degree classification, if their competitor marks, those of their fellow students, were less than they might otherwise have been. Alternatively, they may decide that they could improve their overall marks position in the cohort by such a tactic. Group members may just not fully appreciate the collaborative nature of their group task. On the other hand, they may deliberately choose to ignore it. They may keep some key information to themselves to enhance their own IISA marks to the detriment of the remaining team members assessment marks.

Where there is an IISA component of a GSA module, then it is possible that a group member will not share key findings for the IISA item. Any GSA module design should try to accommodate this

possibility. Such a situation could also occur in a student led IISA module. They may consider this where the overall module marks have a substantial individual component, in the expectation of improving their own marks at the expense of their colleagues.

Researching potential deviant behaviour is challenging. Despite this, it would be an interesting and instructive research topic for future study.

8.2.10 Extraneous, mediating or intervening variables affecting student module marks

The aim of this present study was to explore the relationship between GSA and IISA categories of summative assessment. There are however, far more variables that influence, mediate or otherwise intervene in students module marks than just whether the summative assessment is by group or individual effort. In Entwistle and Wilson (1977), their use of the term, *any single variable,* was important (see chapter 4 introduction). In terms of students module marks, the impact of assessment method may be less important than that of other, unknowable, variables. Entwistle and Wilson also commented on the importance of interaction of variables. It seems clear that there could be large numbers of these. They will be both internal and external to student's studies.

Other often unknown and therefore uncontrollable attributes, which affect a student summative assessment mark include where, when, and how the material to be assessed was taught and where it was learned. For example, the time of day, day of the week, week of term could influence the module mark. So too could whether learning was through private study, mass lecture, seminar or peer group teaching. In addition, was the same teacher and teaching method used throughout the module teaching sessions, or were different methods used.

Similarly, the physical circumstances of the assessment may have an impact on the students' scores. These are the where, when, and how of assessment. For example, materials taught in a series of short lectures could be assessed by equally short multiple-choice individual assessment. Alternatively, the material could be delivered from a weekly schedule of lectures, seminars and laboratory sessions. In this case, the assessment, either GSA or IISA could be at the end of the study year. All these may or may not play a part in the student's assessment mark. In most cases in this present study, the only known variables are the assessment methods and the marks score.

It may not even be too bizarre to suggest that for some students, a contemporary television show

could influence their mark. Most of them would of course have a similar effect on both their GSA and their IISA module marks. The impact could be as much, or even more than, their perception of the assessment topic, and of the method used to assess it. Certainly, the time spent watching the broadcast would detract from their study time. The influence need not all be negative. It could also enhance their understanding of the topic and hence their ability, and attainment in the summative assessment. A long running soap opera may have a negative effect, but a documentary series presented by a specialist in the field may have a positive one. This sort of effect could depend on how the groups were formed. Some of those based on friendship could be distracted by dramas, while some formed from special interest groups could be attracted to a documentary, for example. Such impacts would be unknowable and immeasurable without a very different and more intrusive kind of study. It would need to be appropriately designed, and perhaps require novel calibrated research instruments to be developed and produced.

Bearing in mind the Entwistle and Wilson (1977:5) quotation cited earlier (in the introduction to Chapter 4) unknown variables also include imponderables such as the students' general character, extracurricular interests or personality type. Other examples are family socio-economic background and cultural capital, and the student's level of self-confidence, and wellbeing. These affect both GSA and IISA, to an unknowable extent. They also include the strength and levels of interest in the module subject, and their experience of and anxiety level over the assessment method. These will all covary and may influence any module summative assessment mark. They may vary with time, or with the level of other distractions from academic study.

Although there are studies that look at the link between assessment score and many of these individual variables, none was found that took account of this entire, potentially vast, array of variables. In particular, none was found that tried to control for the impact they may have on students summative assessment marks.

8.2.11 Personality types, learning styles and preferred roles

Another aspect of the effect of GSA on overall marks is that of the interaction of different personality types (see section 3.6.3) and preferred roles of students in their group projects. The individual's role in the group may influence their marks outcome. Their role in the group could either complement or conflict with their own. This could also depend on whether the role was allocated, or was self-selected. Data on students preferred group role was not collected.

An additional assessment complication could arise for groups formed, for instance, to provide a mix of student team role types. Popular forms of these are Belbin, (e.g. Belbin 2004; Aritzeta et al. 2007) and Myers-Briggs (2007). (NB, note the Coffield et al. (2004) reservations about the MBTI method). Assessment would be problematic for both peers and for teachers. If students all assume a different role in the team, there will be no student with the same role in that group with which to compare them. If by chance the group was to consist of students of the same role type then the group is also quite likely to fail, (for example see Belbin (2004), or section 3.6.3.2). Different students in different groups, with the same roles, also seem likely to affect their group outcomes differently because the groups would have different dynamics. There could also be problems properly assessing the same roles across different teams. Peer assessment would just become a guess, based in part on the personal attributes of the group member, e.g. their charisma and personality. It would also depend on the other group member's perception of their peers ability and attainment.

The ethics surrounding Belbin's data collection methods in the 1970's might be more questionable today. His experiments manipulated team attributes. This influenced their outcomes. These students were studying for a high-stakes qualification, the MBA. In section 3.6.3, it was also noted that the concept of learning styles has been discredited by some educationalists.

8.2.12 Stakeholders' perspectives of GSA

8.2.12.1 GSA from the universities' perspective

Some universities may believe that GSA could give them a competitive advantage. There may be several reasons for this:

The first is that it adds to the variety of assessment methods available to teaching staff and module designers. This reduces the overall risk of criticism that their restricted range of teaching and assessment methods was unfair to some students. It may also reduce the need to allocate resources to defend appeals from students who may be unhappy with their degree classifications.

Second, according to some reports, (see section 3.5.1,) employers want graduates to have group working experience. In their graduate recruitment, prospective employers may favour graduates from higher education institutions that use GSA. This in turn would have a beneficial effect on the university's graduate employment statistics, and, when published, may in turn encourage other students to study at their institution. When funding their own employee's university studies,

employers may even prefer universities that include GSA because they also gain further experience in groupwork.

A third reason is that more and more study disciplines and programmes worldwide are embracing GSA as a teaching, learning and assessment method. (See section 4.10.1). Other universities that do not practise it may think that if GSA practice is so popular, then it must be good practice

Fourth, it *may* be cheaper to deliver the module. There could certainly only be one script, video, or poster etc. to mark instead of individual ones for each team member. This may be a major consideration in a minority of disciplines.

8.2.12.2 GSA from the perspective of university teaching staff

University teaching staff experience of GSA may provide them with insights that contradict those of their employer. This is due to their more direct practical experience of it, including how they allocate students to groups. They may also be able to confirm that graduate employers want groupworking experience in their graduate recruits, from their own experience of providing student references to prospective employers. Their networking, as well as research articles on groupwork and GSA would also inform the opinions of teaching staff. They may also have diverging views on whether or not GSA is a more efficient use of resources.

Group working may also be difficult for some modules in terms of the group-dynamics. In such cases, teaching staff may have to devote resources to overseeing the smooth running of the module. In particular, they will have to deal with group dynamics issues including that of the free rider. (See section 3.9.7).

The rationale for using GSA could include either, or both, the future professional benefits for the student and the sheer necessity for teamworking in order to complete a large workload, or learn specialist skills and techniques. Where students understand this, the module may proceed smoothly. Everyone concerned, including teaching staff, may then develop a more positive view of GSA.

8.2.12.3 GSA from the perspective of educational academics

Educationalists may have similar views about GSA to either or both the university and the teaching staff. After all, they may have formal inputs into both of these areas. On the other hand, they may question the validity of using GSA in such a high-stakes public qualification. They may also have

concerns over the defensibility of GSA practice and about its effectiveness in achieving module outcomes. These could also extend to the overall, or module lifetime cost of GSA, compared to other assessment methods. Since the advent of GSA, some academics may suggest the concept of a degree is changing, that a degree is no longer strictly an individual qualification. For them, GSA might be a legitimate assessment method. They may have doubts about sending graduates into careers, without them having had any formal exposure to, or summative assessment in, group-dynamics.

8.2.12.4 GSA from the students' perspective

Some students may already perceive their degree classification to be something of a lottery, see sections 2.7.1 and 3.6.1.2. If they study a GSA module then it seems from this and other studies (Lejk et al. (1999), Hoffman and Rogelberg (2001) and Almond (2009)), that it will probably influence their overall mark. They may have additional concerns because of this, or they may be unaware of it. Their degree would become even more of a lottery than they might first have perceived it to be, because of GSA use. For some, GSA could also add additional stress to assessment experiences that were already stressful. For others it may offer support, reducing their assessment stress. Students will have different expectations about their university experience. They will also have different preferences for the assessment methods.

Student's motivation, general academic ability and interest in the topic will also be an important factor. These marks could be from coursework consisting of either individual or group project essays or oral presentations. They could also be from individual multiple-choice tests, or seen or unseen written examinations. The extent that a university, faculty, or department uses GSA may affect a student's choice of study programme or university. The effect could be either positive or negative. The extent that they and their advisors consider GSA when making their study, discipline, programme and university choices would be an informative future study. It would help HEI policy makers to decide on whether to practice GSA.

8.2.13 Assessing a GSA module by re-sitting

An important part of students summative assessment procedures includes arrangements for their opportunity to re-sit a failed module. The student experience of group working is often part of the rationale for including it in the module teaching, learning and assessment aims. For a GSA re-sit attempt the possibility of an additional GSA item no longer exists. The other members of the group will have moved on. In this case, the re-sit could take the form of, for example, a two-part essay. The candidate could discuss what they learned from the group-dynamics aspect of the project as

well as evaluating their understanding of the topic. The marking scheme of such a module would require careful planning.

8.2.14 Limitations of the present study

The limitations of this present study have been discussed throughout this thesis as the issues arose. This section summarises them.

The background data received for the students from the four sources was inconsistent. For example, the level of detail data received from source D were very different to that envisaged at the conception of the present study. There was no personal background data on the students except their faculty and year of study (see section 6.3). Biographical data could be used to, for example, look for differences in the effect on GSA marking between study disciplines, and student ages, gender and ethnicity. Some previous ability data at programme entry was received in one of the other data sets, but this was incomplete and inconsistent, as was information on the assessment methods from the four data sources. Despite this, the reduced data might not have been the limitation it first seemed. Data as rich in biographical detail as had been originally hoped for may have taken too long for this lone researcher to analyse fully within the scheduled timescale of the present study.

The generalizability of the study findings cannot be assumed from this present study. A wider ranging further study will be needed to determine whether there is a typical marks distribution pattern. Establishing this IISA/GSA marks distribution pattern for any module will be problematical. The first issue would be to define the student population. For example, would this be all HEIs that practice GSA in the world, or only those that taught in English? An additional issue is how to deal with varying students abilities. In the UK, for example, students of different academic ability levels could choose to study at different kinds of university, e.g. Oxbridge, Russell Group, 1994 Group, a modern university or a university created from the post 1992 polytechnic divide, and this would have to be accommodated. The extent to which the university practices GSA may also contribute to students choices of study institution. It may not be possible to identify a typical IISA/GSA marks distribution pattern, just as it may not be possible to identify a typical student, because there are too many variables. A typical distribution from data with one set of variable values may be different to that from data with different variable values.

In systematic review meta-analysis, the differences in methods, e.g. interventions and sampling

techniques, between studies, are labelled moderator variables. (See Lipsey (2003) and section 7.6). The data that was used in this present study were collected from four different sources and were at different levels of detail. In each of them, the moderator variables would have been context specific and therefore different for each module, programme, department and faculty from whence the data originated. In the data from source D, for example, the IISA marks were the total module mark for each of the non-GSA designated modules that the student studied. The GSA mark was from at least one GSA item mark. The marks that contributed to the GSA module marks also included IISA assessment item marks, to an unknown extent. The different data sources: A, B, C and D, did not share a common marking scheme, but GSA was common to all of them.

If the sampling frame had been broader and four years of data had been collected from the source of data D instead of only three, then it would have included at least one complete cohort of data from each of the four-year fully integrated Masters programmes. The engineering programmes from the university in the northeast of England, from which the D data was collected, for example, included several GSA modules, at all levels of their MSci programme study years. If, additionally, Bachelors and Masters first-degrees had then been differentiated, this would have allowed the possibility of comparing the impact of GSA marking on students studying four-year first-degree programmes, compared to those whose studies were for three years. The existence of anything other than a three-year (full-time) first degree, including durations of less than three-years, was not considered during the planning of this present study. It was an important oversight because it also excluded the possibility of collecting data from specialist, professional, intensive, shorter duration, and conversion programmes.

An additional limitation of the present study was the lack of information on the follow up of the students, i.e. to what extent did GSA (and IISA) scores predict students later career success. A simple postal survey could have dealt with a ‘first destination’ analysis, although career progression could be more problematic to study.

This present study included data from as many sources as were available, despite their lack of a common marking scheme, biographical data etc. An alternative could have been to analyse data from only one source, e.g. source D. Including all four data sources in the analysis was more representative of the global student population. Additionally it would, of course have been possible to reduce data from source C to the level of detail in data source D. That would have just wasted

the richness of detail generously offered for this present study. The issue of the hierarchical nature of the data is raised in section 5.2.

The greatest limitation of the multilevel modelling in section 7.5 of this present study arose due to the limited sources of the data that were available for it. A prerequisite of MLM is that the data needs to be randomly sampled from the population. This is problematic because the data sets collected were a convenience sample, from only four sources rather than from a random sample of the population, that were available at the time of the present study. Although they were a convenience sample, they were not, *"subjectively determined by the investigator"* (LEMMA 2009: Module 1).

8.2.15 Human error and other alternative reasons for the IISA and GSA data difference

For the most part, the Regression charts in Appendix 14 and Appendix 12 reveal similar data distributions to that of the pilot study, (shown in Figure 2).

One of the reasons for the difference between the IISA and GSA marks data could simply be human error in the present study. This seems unlikely because it would have required widespread and systematic, rather than random data corruption. Additionally, as mentioned earlier, the data were crosschecked before, during and after processing, the data requirements having been negotiated with and explained to three separate gatekeepers, the fourth being already in the public domain, see Knight (2004).

It also seems probable that GSA may not operationalize student attainment the same way in every GSA application. There are several forms of IISA, e.g. written examination, coursework, viva etc. Similarly, the GSA method may only be useful as a generic label of GSA to differentiate it from IISA. There may be little or no consistency between practitioner methods. Like all assessment, they are context specific. This means that for the most part, they would be different to each other. These may include methods of marking only GSA products, or the group process, or both. The modules could also either be with or without an IISA report item or a traditional written IISA examination. This means that for example, the marking schemes of the data from source A modules may have been different between data subsets A1 to A5. It would probably be different again, from the GSA marking methods of the other data sets. GSA may attempt to measure several covarying learning concepts. They could include the students understanding of the topic,

as well as their attainment in key groupworking skills. (Also, see section 3.5.2.)

8.2.16 Conclusions

This work built on earlier work of Lejk et al. (1999), Hoffman and Rogelberg (2001) and Almond (2009) and, to a slightly lesser extent, Simonite, e.g. 2001b. It has added to the body of knowledge of the impact of summative assessment of student academic group work because, unlike earlier studies, the data were of a large number of data subjects from a variety of HEI study modules. This greatly improved the generalisability of the finding.

The main conclusion from this present study is that GSA is unfair. There were systematic differences between students' IISA and GSA marks (see for example section 7.4.4). In this present study, the extent and direction of the systematic marks differences varied between students of high and low ability. They also varied between faculties. In addition, they also varied between levels of study, i.e. between undergraduate and post-graduate programmes, and/or between HEIs and/or countries.

The null hypothesis H_0, that the two summative assessment methods have the same impact on students overall marks (Table 1), is rejected. The alternative hypotheses H_1, that they have a different impact on students overall marks, and H_2, that uncorrelated data will show the same pattern in a dual-line regression chart as the present study data, were supported. Although the modules correlations did not support the alternative hypothesis H_3, that there will be a low level of correlation between IISA and GSA marks, neither did they disprove it (see for example section 7.3).

The findings of this present study were supported, separately, by those of Lejk et al. (1999), Hoffman and Rogelberg (2001) and Almond (2009) although it is difficult to compare them directly. The marks variation differences found between students of high and low ability was supported by Lejk et al. (1999).

Hoffman and Rogelberg (2001) studied students preferred project group grading procedures. They found that students with a high grade point average (GPA) would avoid GSA project work *where they all get the same grade*, while students with a low GPA would embrace it. From the evidence of this present study, this seems likely to be because their scores would be higher because of GSA than they would have been, based on their IISA results. On the other hand, results from this present study predict that GSA will reduce higher ability students' marks. They may be more

reluctant to engage with GSA items because of this.

This present study upheld the working hypothesis that the GSA concept may be unreliable. During the literature searches and reviews however, it became apparent that GSA is no more problematic than, for example, an essay assessment or an unseen written, timed examination, (see for example, section 4.6.1). Although small and self-funded, with data collected from a convenience sample, this study has shown, as mentioned earlier, that in the data available there was a systematic difference between the IISA and the GSA marks. It has also shown that the extent and direction of the differences varied between students of high and low ability, and between faculties.

A broader study of the use of GSA is necessary to assess the generalizability of the findings. More data from a wider ranging study would also make outlier recognition easier by making data pattern identification easier, (see section 7.1).

8.2.17 Recommendations for practice

8.2.17.1 Cease GSA practice

Group working and group summative assessment are conceptually different practices. The former is potentially very useful for students. Firstly as an alternative method of learning and teaching, and secondly in providing group dynamics experience, which may be useful for their future professional careers. On the other hand, GSA practice should cease in its present form. There should be no general migration to it, on the rather doubtful grounds of efficient assessment, without further supporting research evidence of its effect on overall marks. In particular, the advantages and disadvantages should be clearly understood by all stakeholders, which would surely lead to a radical change to current GSA practice. Several potential ways forward are presented below.

8.2.17.2 Include GSA in a non-contributing study year

The issue of the unfairness of GSA marking effects on final degree classifications is easy to eliminate. The solution is to restrict GSA use to those modules where the marks do not count towards it. This could be where GSA was used exclusively in a non-contributing first year of programme study. In many HEIs, for example where all study levels currently contribute towards the final degree classification, this may require changes to their regulations. Its use could also be restricted to a non-contributory, common foundation year following the model suggested by the Candy et al. *excessive specialisation* solution (1994:117).

8.2.17.3 Separate IISA and GSA modules

A less andragogically satisfactory though nonetheless practical solution would be to have two separate modules on the same topic in a programme with one module assessed individually and the other module by the GSA method. There are two main disadvantages with this model. The first is that in such a programme, two modules could cover similar ground, although questions would be set on different aspects of the topic. Second, this less efficient duplication of effort also bears an opportunity cost in that there would be less room on the programme for other topics.

In this model, the IISA module mark would contribute to the degree classification but the GSA module mark would not. This might also require a change to the HEI degree regulations. For study progression and for the IISA mark to count towards the overall mark, the student would also have to pass the GSA module at the threshold level. (See sections 7.2.4 for additional comment on threshold marking.) Alternatively, every group member of a *non-contributing* GSA module could receive the same mark. This would not overcome the problem of fairness in the assignment of students to groups. It would however make it less of a high stakes issue because it would eliminate the unfairness of the GSA mark on contributing individual overall marks.

8.2.17.4 Include a JumpStart style option in the study programme

The Southampton University *Jumpstart* model (see section 3.12) would seem to be an ideal method of introducing students to group dynamics voluntarily, without detriment to their overall mark and subsequent degree classification. The experience would also be of interest to most future employers and it could be included in the student transcript. There may however be issues of group membership allocation where it was used in a core module, especially at an early study level, e.g. level 1. In addition, in such an event, it would need to be summatively assessed or it would not be treated seriously, (see section 4.5.6). It would also need a common activity for perhaps several thousands of new students. This could be in the form of a group presentation on a generic topic on, for example, how to conduct literature searches, or simply a group project about group-dynamics. One of the main disadvantages of this solution might be the effect that the marking requirement might have on the academic markers for such a very large cohort. The most obvious disadvantage to this option is that, unlike the *Jumpstart* example, participation would no longer be voluntary.

8.2.17.5 Assess only the individual components of group project modules

Another solution is only to summatively assess the independent individual assessment components of the group project module. This is an alternative to the section 8.2.17.3 described earlier. It has

the advantage of keeping the same number of modules in the programme and not increasing either the staff or student workload. The group project item could receive a threshold mark *P* (see next section), be left unassessed, or be formatively assessed. This would overcome the *degrees are awarded to individuals, not groups* conundrum. The group project could be, for instance, laboratory work, or a project report. It could be a group project, with the report submitted individually. A further alternative is an individual essay, written recommendations, or oral presentation, on the group-dynamics aspect of the group task. The task could be, for example, to redesign the group task so it becomes possible for an individual to complete it. Another possibility is to use GSA for a formative assessment with an individual summative assessment item submission. This would follow the formative feedback. The drawback of such a scheme is, as discussed earlier, that it may not be treated as seriously as a task that is summatively assessed.

8.2.17.6 Use a threshold mark for all GSA modules marks

A novel assessment item, such as a collaborative group project, would seem to require a novel summative assessment solution. The present study data, from source D, included data from 44 students who received a module mark indicated by the letter P, (also, see section 7.2.4). Oral evidence from senior academics has also confirmed use of this threshold P mark method. This meant that the students concerned received a contributory pass grade, rather than an actual score of between 0 and 100, (see D data set distribution in Figure 21). These data were for modules where the student spent a year either studying at an HEI abroad, or working for an employer outside the university, either in the UK or abroad. This designation could be used to mark GSA modules.

8.2.17.7 Promulgate better information on GSA methods and findings

All practitioners of GSA should strive to make the assessment method as transparent as possible to all stakeholders, to whom they should promulgate the method. Where appropriate they should include their results and findings. This would mean that, over time, GSA would become a safer practice. Only when stakeholders collectively understand the concept of GSA, will it be a valid assessment method. Currently, it is possible that for most non-specialist stakeholders, one degree assessed by one method by one university, could seem very much like any other. Under such circumstances, they could place similar values on similar degree classifications from different universities and assessment regimens.

8.2.18 Recommendations for, and questions suggesting, further study

This section presents recommendations for, and questions suggesting, further study. If further

studies support the findings from this one, then stakeholders' confidence in the reliability and validity of GSA practice should improve.

There is a wide range of issues concerning GSA practice that have arisen during this present study, which remain unanswered. For example how important is an understanding of the programme assessment methods that may be available to universities, for students to be able to make an informed choice of their preferred study university? Alternatively, to what extent is a complete understanding of their implications a prerequisite for first-degree study?

Underpinning the recommendations and questions in this section is the issue of how GSA compares to other summative assessment methods, e.g. essays or timed, unseen, written examinations, or viva voce. Reference material directly relevant to this present study was scarce.

One of the *Standard's* (Baker et al. 1999), principal definitions of fairness was *equitable treatment in the testing process* (see section 4.9.2). This clause could disqualify GSA because of the methods of assigning students to groups. Publishing further research results will help to make the fairness of GSA more transparent to all stakeholders.

Experience of collaborative working in paid employment could help or hinder students studies in terms of developing their interpersonal skills. It could affect how they engage with group projects. It would be useful to explore the difference in GSA marks between those students who have experience of paid employment compared to those who do not. The results could help prospective students to decide whether to head straight to university from school, or first to spend time in paid employment to gain some team working experience. They could even have an impact on university admissions policy.

A much longer-term problem with GSA is how do the results from it relate to graduate success in later life? Problems with such a study would include how to define success, difficulty with tracking and maintaining contact with graduates over an extended study, and issues of the reliability of the self-reported data. A careful study design could accommodate these difficulties, as has clearly been achieved by, for example, the British Household Panel Survey (BHPS 2011) or Tymms (1995).

Other questions arising from this present study include for example:

Is GSA misused or overused by some HEIs and used properly or underused by others?

To what extent is the effect of GSA similar across all levels of HE taught programmes, i.e. at both undergraduate and postgraduate level. (Also, see section 6.1.)

To what extent are the assumptions behind the rationale for GSA practice realistic? GSA may not always be a sure shortcut to greater efficiency and improved profitability for the institution that practices it. What is the true lifetime cost of GSA practice in HEIs for the various groups of stakeholders, e.g. the HEI and their staff, as well as students?

How does including GSA in an institution's module learning and teaching scheme impact on student registration, and on their dropout and failure rates in comparison to those HEIs that do not practice GSA?

How does student age and/or study level impact on their GSA marks?

How does the GSA effect differ between disciplines, schools or departments and faculties? Why are there differences? Could differences be expected?

Are some types of assessment methods more reliable than others, for example, how does the type of GSA, e.g. an essay, a written or oral report, or an artefact, video, Wiki, or poster, impact on overall marks?

To what extent can groupwork be learned by students and how does the practice effect influence their results?

From a philosophical or assessment policy point of view, was student GSA just a bold and brave educational experiment, to expose students to high stakes group working and to try to educate them in interpersonal skills, and is it now time for radical change to practice?

"Who, then, shall conduct education so that humanity may improve?" (Dewey 1916:122)

References

Ackerman, A and S Plummer. (2004, downloaded 30 March 2004). "Examination into the use, place and efficacy of group work in university courses: a work in progress report of a current research project." Retrieved 30 March 2004, from http://www.aare.edu.au/94pap/ackea94.306

Alford, R R and R Friedland (1985). Powers of theory: capitalism, the state, and democracy. Cambridge, Cambridge U P.

Alkin, M C, Ed. (2004a). Evaluation Roots: tracing theorists' views and influences. Thousand Oaks, Sage.

Alkin, M C (2004b). Comparing evaluation points of view. Evaluation Roots: tracing theorists' views and influences. M. C. Alkin. Thousand Oaks, Sage**:** 3-11.

Alkin, M C (2004c). Context-adapted utilization. Evaluation Roots: tracing theorists' views and influences. M. C. Alkin. Thousand Oaks, Sage**:** 293-303.

Allen, J and R Lloyd-Jones (2006). The Assessment of Group Work and Presentations in the Humanities: A guidebook for tutors., Sheffield Hallam University.

Almond, R J (2006). The Effect of Group Summative Assessment Marking on Student Marks. (Unpublished MA Dissertation), Durham University School of Education.

Almond, R J (2009). "Group Assessment: comparing group and individual undergraduate module marks." Assessment & Evaluation in Higher Education 34(2): 141-148.

Amato, C H and L H Amato (2005). "Enhancing student team effectiveness: Application of Myers-Briggs Personality Assessment in business courses." Journal of Marketing Education 27(1): 41-51.

Anderson, D F. (1953). "Tests of Achievement in English Language." Retrieved 04 March, 2008, from http://eltj.oxfordjournals.org/.

Archambault, R D, Ed. (1965). Philosophical analysis and education. London, Routledge & Kegan Paul.

Aritzeta, A, S Swailes and B Senior (2007). "Belbin's Team Role Model: Development, Validity and Applications for Team Building." Journal of Management Studies 44(1): 96-118.

Atkins, M (1995). What Should We Be Assessing? Assessment for learning in higher education. P. Knight. London, Kogan Page.

Bacon, D R (2005). "The Effect of Group Projects on Content-Related Learning." Journal of Management Education 29(2): 248-267.

Bacon, D R, K A Stewart and W S Silver (1999). "Lessons from the best and worst student team experiences: How a teacher can make the difference." Journal of Management Education 23(5): 467-488.

Bacon, D R, K A Stewart and S Stewart-Belle (1998). "Exploring Predictors of Student Team Project Performance." Journal of Marketing Education 20(1): 63-71.

Baird, J-A (1998). "What's in a name? Experiments with blind marking in A-level examinations." Educational Research 40(2): 191-202.

Baker, E, P Sackett, L Bond, L Feldt, D Goh, B Green, E Haertel, J-I Hansen, S Johnson-Lewis, S Lane, J Materazzo, M Meier, P Moss, E Olmedo and D Pullin (1999). Standards for educational and psychological testing. Washington DC, American Educational Research Association.

Barfield, R L (2003). "Students' perceptions of and satisfaction with group grades and the group experience in the college classroom." Assessment & Evaluation in Higher Education 28(4): 355-369.

Barr, T F, A L Dixon and J B Gassenheimer (2005). "Exploring the "Lone Wolf" Phenomenon in Student Teams." Journal of Marketing Education 27(1): 81-90.

Barrett, H. (2004). "Using Technology to Support Alternative Assessment and Electronic Portfolios " Retrieved March, 2004.

Barton, M, C Davis, B Harris, S Hibberd, M Jenkins, T Katz and E Wilcock (2002). LTSN Engineering working group report: Assessment of individuals in teams.

Baty, P (2009). 'Massification' takes toll on professoriate standards. Times Higher Education.

Baume, D (2001). A Briefing on Assessment Portfolios, LTSN Generic Centre November 2001. Assessment Series No.6.

BCS. (2007). "Guidelines on course registration." Retrieved March, 2008, from http://www.bcs.org/accreditation

Becker, H S (1995). Making the Grade Revisited. Making the Grade (Reprint). Becker, Geer and Hughes. New York, Wiley.

Becker, H S, B Geer and E C Hughes (1968). Definition of the Situation: Organization rules and the importance of grades. New York, Wiley.

Belbin, R M (2004). Management Teams, Why They Succeed or Fail, Second Edition, Butterworth-Heinemann.

Bennett, N, E Dunne and C Carre (2000). Skills development in higher education and employment.

Buckingham, SRHE and Open University Press.
BHPS. (2011). "British Household Panel Survey." Retrieved 02 September 2011, from http://www.esds.ac.uk.
Bion, W R (1961). Experiences in Groups, Tavistock Publications.
Black, P (1998). Testing: Friend or Foe?, Falmer Press.
Borsboom, D, G J Mellenbergh and J v Heerden (2004). "The Concept of Validity." Psychological Review 111(4): 1016-1071.
Boud, D (1990). "Assessment and the Promotion of Academic Values." Studies in Higher Education 15(1): 101-111.
Boud, D (1995b). Assessment and learning: contradictory or complementary? Assessment for Learning in Higher Education. P. Knight. London, Kogan Page: 35-48.
Boud, D (1998). Assessment and learning - unlearning bad habits of assessment. Effective Assessment at University. University of Queensland.
Boud, D (2000). "Sustainable assessment: rethinking assessment for the learning society." Studies in Continuing Education 22(2): 151-167.
Boud, D, R Cohen and J Sampson (1999). "Peer Learning and Assessment." Assessment & Evaluation in Higher Education 24(4): 413-426.
Boud, D, R Cohen and J Sampson, Eds. (2001). Peer Learning in Higher Education. London, Kogan Page.
Bowling, A (2002). Research Methods in Health 2nd Edition. Buckingham, Open University Press.
Boyne, P R (2007). Module Convenor. R. J. Almond.
Brennan, D J (2008). "University student anonymity in the summative assessment of written work." Higher Education Research & Development 27(1): 43-54.
Brew, A (1999). Towards autonomous assessment. Assessment Matters in Higher Education. S. Brown and A. Glasner. Buckingham, OU Press.
Brown, C A and K McIlroy (2011). "Group work in healthcare students' education: what do we think we are doing?" Assessment & Evaluation in Higher Education 36(6): 687-699.
Brown, G (2001). Assessment: A guide for lecturers, LTSN Generic Centre.
Brown, G, J Bull and M Pendlebury (1997a). Assessing Student Learning in Higher Education. London, Routledge.
Brown, R W (1995). Autorating: Getting individual marks from team marks and enhancing teamwork. Frontiers in Education, Pitsburgh.
Brown, S (1999a). Institutional strategies for assessment. Assessment matters in higher education: choosing and using diverse approaches. S. Brown and A. Glasner. Buckingham, OU Press.
Brown, S, V Butcher, L Drew, L Elton, L Harvey, P Kneale, P Knight, B Little, N Moreland, M Yorke, R Bhanot, B Chalkley, G Crust, P Elkes, J Gawthrope, S Harbour, J Jones, D Macfarlane-Dick, T Overton, L Philips, J Terry, K Trehan and M Willis (2006). Learning & Employability, 8. Pedagogy for Employability. York, The Higher Education Academy (LTSN).
Brown, S and P Knight (1994). Assessing Learners in Higher Education. London, Kogan Page.
Brown, S, P Race and B Smith (1996). An Assessment Manifesto. 500 Tips on Assessment, Kogan Page.
Bryman, A (2004). Social Research Methods. Oxford, Oxford University Press.
Burd, E, S Drummond and B Hodgson (2003). Using Peer & Self Assessment for Group Work. 4th Annual LTSN-ICS Conference, NUI, Galway, LTSN Centre for Information and Computer Sciences.
Burgess, R (2007). Beyond the honours degree classification: The Burgess Group final report, Universities UK.
Calkins, L B (2004). Enron Fraud Trial Ends in 5 Convictions. Washington Post.
Camara, W J and S Lane (2006). "A Historical Perspective and Current Views on the Standards for Educational and Psychological Testing." Educational Measurement 25(3): 35-41.
Candy, P C (1995). Developing lifelong learners through undergraduate education. A Focus on Learning; Proceedings of the 4th Annual Teaching Learning Forum, Edith Cowan University, February 1995. L. Summers (Ed). Perth, Edith Cowan University. http://lsn.curtin.edu.au/tlf/tlf1995/candy.htm, accessed 03/06/2004: ii-viii.
Candy, P C, G Crebert and J O'Leary (1994). Developing lifelong learners through undergraduate education, Australian Government Publishing Service: 348.
Cheng, W and M Warren (2000). "Making a Difference." Teaching in Higher Education 5(2): 243-255.
CIEA, C I o E A. (2009). "Purposes of Assessment." Retrieved 08 April 2009, from http://www.ciea.org.uk/knowledge_centre/articles_speeches/general_articles/purposes_of_assessment.aspx.
Clark, A (2010). Jeffrey Skilling, convicted Enron boss, claims trial was prejudiced. The Guardian.
Cleaver, C S A (1995). RSA Inquiry: Tomorrow's Company.
Clegg, F (2002). Simple Statistics, 18th ed, Cambridge University Press.

Coe, R (2002). It's the Effect size, Stupid: What effect size is and why it is important. British Educational Research Association, University of Exeter, Education-Line.
Coe, R (2004). Issues arising from the use of effect sizes in analysing and reporting research. But what does it mean? I. Schagen and K. Elliot. Slough, National foundation for educational research**:** 80-100.
Coe, R (2010). Learning Styles: Research & Fashion, CEM Durham University.
Coffield, F, D Moseley, E Hall and K Ecclestone (2004). Should we be using learning styles? London, Learning and Skills Research Centre.
Cohen, M and A Mullender (1993). Gender and Groupwork, Routledge.
Conway, R, D Kember, A Sivan and M Wu (1993). "Peer Assessment of an Individual's Contribution to a Group Project." Assessment & Evaluation in Higher Education 18(1): 45-56.
Cook, A (2001). "Assessing the Use of Flexible Assessment." Assessment & Evaluation in Higher Education 26(6): 539-549.
Cox, R (1967b). "Resistance to change in examining." Universities Quarterly 21: 352-358.
Crabtree, M (2009). Me and You and Everyone We know: PG workshop 01/07/09.
Craib, I (1992). Modern Social Theory: from Parsons to Habermas (2nd Ed.). London, Harvester Wheatsheaf.
Crebert, G. (2007). "A snapshot of generic skills development at Griffith University." Retrieved 03/01/2007, from http://www.gu.edu.au/centre/griffith_graduate/snapshot_gu.pdf.
Creswell, J W (2003). Research Design: Qualitative, Quantitative, and Mixed Methods Approaches. Thousand Oaks, Sage.
Cronbach, L J (1970). Essentials of psychological testing (Third Ed.). New York, Harper.
Cronbach, L J and P E Meehl (1955). "Construct Validity in Psychological Tests." Psychological Bulletin 52(4): 281302.
Crotty, M (1998). The Foundations of Social Research. London, Sage.
Danili, E and N Reid (2005). "Assessment formats: do they make a difference?" Chemistry Education Research and Practice 6(4): 204-212.
Dearing. (1997). "The national committee of enquiry into higher education." Retrieved July 2004, 2004, from http://www.leeds.ac.uk/educol/ncihe/ 25/June 2009.
Dedrick, R F, J M Ferron, M R Hess, K Y Hogarty, J D Kromrey, T R Lang, J D Niles and R S Lee (2009). "Multilevel Modeling: A Review of Methodological Issues and Applications." Review of Educational Research 79(1): 69-102.
Dee, M (2008). Young People, Public Space and Citizenship, Unpublished Doctoral Thesis. Brisbane, Queensland University of Technology.
Delandshere, G (2001). "Implicit Theories, Unexamined Assumptions and the Status Quo of Educational Assessment." Assessment in Education 8(2): 113-133.
Descriptors. (2008). "University of Durham undergraduate qualification descriptors." Retrieved 31 12 2008, from http://www.dur.ac.uk/resources/university.calendar/volumeii/2007.2008/ugqualdes.pdf.
DeVita, G (2001). "The Use of Group Work in Large and Diverse Business Management Classes." The International Journal of Management Education 1(2): 26-34.
Dewey, J (1916). Democracy and Education, Rector and Visitors of the University of Virginia.
Dewey, J (1938). Logic: The theory of inquiry. New York, Henry Holt and Company.
Dore, R P (1997). The diploma disease: education, qualification and development, second edition. London, Institute of Education, University of London.
Downie, N (2001). "Assessing Group work in Higher Education." Innovation 2001, Learning & Teaching Journal, Nottingham Trent University(5).
Drury, H, J Kay and W Losberg (2003). Student satisfaction with groupwork in undergraduate computer science: do things get better? Australasian Computing Education Conference (ACE2003), Adelaide, Australia.
DU_2009Brochure (2009). Creating the Future, Durham University.
Duff, A and T Duffy (2002). "Psychometric properties of Honey & Mumford's Learning Styles Questionnaire." Personality and Individual Differences 33: 147-163.
Durham University. (2006). "Modules By Department." Retrieved 11/07/2006, from http://www.dur.ac.uk/faculty.handbook/listdeptmodules/.
Earl, S E (1986). "Staff and Peer Assessment- Measuring an Individual's Contribution to Group Performance." Assessment & Evaluation in Higher Education 11(1): 60-69.
Ebel, R L (1961). "Must all tests be valid?" American Psychologist 16(10): 640-647.
Elander, J (2004). "Student assessment from a psychological perspective." Psychology Learning and Teaching, 3(2): 34-41.
Elton, L (2004). "A Challenge to Established Assessment Practice." Higher Education Quarterly 58(1): 43-62.
Elton, L and B Johnston (2002). Assessment in Universities: a critical review of research, Higher Education Academy (formerly Learning and Teaching Support Network).

Entwistle and Wilson (1977). Degrees of excellence: the academic achievement game. London, Hodder & Stoughton.
Entwistle, N (1996). Recent research on student learning and the learning environment. The Management Of Independent Learning. Tait and Knight. London, Kogan Page: 97-112.
Erwin, T D (2000) The NPEC Sourcebook on Assessment, Volume 1 N. C. f. E. S. U.S. Department of Education
Evans, C A (2003). The Relationship Between the Cognitive Style(s) and Preferred Teacher Style(s) of PGCE Students. School of Education, Durham. Doctor of Education: 265.
Falchikov, N (1986). "Product Comparisons and Process Benefits of Collaborative Peer Group and Self Assessments." Assessment & Evaluation in Higher Education 11(2): 146-166.
Farrell, M J and N Gilbert (1960). "A type of bias in marking examination scripts." British Journal of Educational Psychology 30(1): 47-52.
Fellenz, M R (2006). "Towards Fairness in Assessing Student Groupwork." Journal of Management Education 30(4): 570-591.
Field, A (2009). Discovering Statistics Using SPSS (Third Edition). London, Sage.
Fletcher, J (2007). "What is heterogeneity and is it important?" British Medical Journal 334: 94-96.
Ford, L (2007). Graduates lacking soft skills, employers warn. Education Guardian.
Forrest-Presley, D L, G E MacKinnon and T G Waller, Eds. (1985). Metacognition, Cognition, and Human Performance. London, Academic Press.
Forsyth, D R (2006). Group Dynamics, international student edition. Belmont, CA, Thomson Wadsworth.
Gale, K, K Martin and G McQueen (2002). "Triadic Assessment." Assessment & Evaluation in Higher Education 27(6): 557-567.
Gammie, E and M Matson (2007). "Group Assessment at Final Degree Level." Accounting Education 16(2): 185-206.
Garland, D (1996). Using research to improve student learning in small groups. Improving student learning: using research to improve student learning. G. Gibbs. Oxford, Oxford Brookes: 224-230.
Gatfield, T (1999). "Examining Student Satisfaction with Group Projects and Peer Assessment." Assessment & Evaluation in Higher Education 24(4): 365-377.
Gersick, C J G (1989). "Marking time: Predictable Transitions in Task Groups." Academy of Management Journal 32(2): 274-309.
Gersick, C J G (1990a). Ch 5 The Students. Groups That Work (and Those That Don't) Creating Conditions for Effective Teamwork. J. R. Hackman. San Francisco, Jossey Bass.
Gessner, R, Ed. (1956). The Democratic Man: Selected writings of Eduard Lindeman. Boston, Beacon Press.
Gibbs, Habashaw and Habeshaw (1988). 53 Interesting Ways To Assess Your Students. Bristol, Technical and Educational Services Ltd.
Gibbs, G, Ed. (1994b). Improving Student Learning. Oxford, The Oxford Centre for Staff Development.
Gibbs, G (1995a). Assessing student centred courses, Oxford Centre for Staff Development, Oxford Brookes University.
Gibbs, G (1995b). Learning in Teams: A tutor guide. Oxford, The Oxford Centre for Staff Development.
Gibbs, G (1999). Using Assessment Strategically To Change The Way Students Learn. Assessment Matters In Higher Education: Choosing and using diverse approaches. S. Brown and A. Glasner. Buckingham, The Society for Research into Higher Education & Open University Press: 41-53.
Gibbs, G. (2009, 31 July 2011). "The assessment of group work: lessons from the literature." from http://www.brookes.ac.uk/aske/documents/Brookes%20groupwork%20Gibbs%20Dec%2009.pdf.
Gidman, W, D Wright and K Marsden (2008). "Team-working styes of first-year pharmacy students." The International Journal of Pharmacy Practice(Supplement 1): A10.
Gielen, S, F Dochy and P Onghena (2011). "An inventory of peer assessment diversity." Assessment & Evaluation in Higher Education 36(2): 137-155.
Gipps, C V (1994). Beyond Testing: Towards a Theory of Educational Assessment. London, Falmer.
Glaser, R (1963). "Instructional Technology and the Measurement of Learning Outcomes: Some Questions." Americal Psychologist 18: 519-521.
Glass, G V, P D Peckham and J R Sanders (1972). "Consequences of Failure to Meet Assumptions Underlying the Fixed Effects Analyses of Variance and Covariance." Review of Educational Research 42: 237-288.
Goldfinch, J (1994). "Further Developments in Peer Assessment of Group Projects." Assessment & Evaluation in Higher Education 19(1): 29-35.
Goldfinch, J and R Raeside (1990). "Development of a Peer Assessment Technique for obtaining

Individual Marks on a Group Project." Assessment & Evaluation in Higher Education 15(3): 210-231.
Grajczonek, J (2009). The good, the bad, the ugly. ATN (Australian Technology Network) Assessment in Different Dimensions, Assessment Conference 2009, Melbourne.
Gray, D E (2004). Doing research in the Real World, Sage.
Greenan, K and S Alexander (2003). Enhancing group working and peer assessment through a virtual learning environment. Business education support team (BEST) conference, Brighton.
Griffiths, M (2002). Marked problems 2. Times Higher Education.
Griffiths, M (2008). Slide Effect. R. J. Almond.
Hackman, J R. (2004). "Leading teams." Retrieved 11 April 2008, from http://www.emeraldinsight.com/Insight/ViewContentServlet?Filename=Published/EmeraldFullTextArticle/Articles/1350100306.html
Hamerton, P G (1887). The intellectual life. New York, Macmillan and Co.
Hammersley, M (2003). Too good to be false? British Educational Research Association Annual Conference, Heriot-Watt University, Edinburgh.
Hardin, R. (2003). "The Free Rider Problem." Retrieved 08/September/2006, from http://plato.stanford.edu/archives/sum2003/entries/free-rider/.
Harpe, B d l, A Radloff and J Wyber (2000). "Quality and Generic (Professional) Skills." Quality in Higher Education 6(3): 231-243.
Hartley, J (1998). Learning and studying: a research perspective. London, Routledge.
Hartog, P and E C Rhodes. (1935). "International Institute Examinations Enquiry: An Examination of Examinations, review." Retrieved 21 February, 2008.
Harvey, L, S Moon and V Geall. (1997). "Graduates work: Organisational change and students' attributes." Retrieved 03/06, 2004, from http://www.uce.ac.uk/crq/publications/gw/gwcon.html
Haywood, J (2000). Assessment in higher education: student learning, teaching, programmes and institutions. London, Kingsley.
HEA (2006). Learning & Employability, series one, Pedagogy for employability. P. M. Yorke, Higher Education Academy Pedagogy for Employability Group. 2006.
Heathfield, M (1999). How to assess student groupwork. Times Higher Education
Helsinki, U o. (2004). "Evaluation of the quality of education and the degree programmes of the University of Helsinki." Retrieved 15 July 2004, from http://www.helsinki.fi/opintoasiainosasto/opintojen_kehittamisyksikko/panelreport_medi_dental.html
Heron, J (1988). Assessment Revisited. Developing Student Autonomy in Learning, Second Edition. Boud. London, Kogan Page**:** 77-90.
Heywood, J (1989). Assessment in higher education 2nd Edn. Chichester, Wiley.
Hill, B J (1972). "An Investigation into the Consistency of Marking Examination Scripts in B.Sc. Part I in Mechanical Engineering." Higher Education 1(2): 221-227.
Hoffman, J R and S G Rogelberg (2001). "All Together Now? College Students' Preferred Project Group Grading Procedures." Group Dynamics: Theory, Research, and Practice 5(1): 33-40.
Holmes, L (1998). One more time, transferable skills don't exist ... (and what we should do about it). Higher Education for Capability, 'Embedding Key Skills Across the Curriculum', Nene College, Northampton, 27 February 1998.
Hornby, W (2003a). "Assessing Using Grade-related Criteria: a single currency for universities?" Assessment & Evaluation in Higher Education 28(4): 435 - 454.
Hornby, W (2003b). "Maintaining Standards? An Analysis of Assessment Practices in a Business School." The International Journal of Management Education 3(3): 1-17.
Horowitz, H L (1986). "The 1960s and the Transformation of Campus Cultures." History of Education Quarterly 26(1): 1-38.
Huanyin, Y (1993). "Confucius (K'ung Tzu) 551-479 BC)." Prospects: The Quarterly Review of Comparative Education XXIII(1/2): 211-219.
Huot, B (1996). "Towards a new theory of writing assessment." College Composition and Communication (CCC) 47(4): 549-566.
Huxham, M and R Land (2000). "Assigning students in group work projects." Innovations in Education and Teaching International 37(1): 17-22.
Hyland, T (1997). The skills that fail to travel. The Times Higher Education Supplement.
Isaacs, G (2002). Assessing Group Tasks, Teaching & Educational Development Institute, University of Queensland.
James, R, C McInnis and M Devlin. (2002). "Assessing Learning in Australian Universities." Retrieved 04 August, 2007, from http://www.cshe.unimelb.edu.au/assessinglearning/docs/AssessingLearning.pdf.
Janis, I L (1972). Victims of groupthink. Boston, Houghton Mifflin.

Jaques, D (2000). Learning in groups, Kogan Page.
Jaques, D and G Salmon (2007). Learning in Groups, Routledge.
JISCinfoNet. (2007). "What do we mean by assessment?" Retrieved 07 06 2007, from http://www.jiscinfonet.ac.uk/Infokits/effective-use-of-VLEs/e-assessment/assess-purpose.
Johnson, C W. (2005). "What are Emergent Properties and How Do They Affect the Engineering of Complex Systems? ." Retrieved 01 November 2009.
Joyce, H. (2001). "Adam Smith and the invisible hand." Retrieved 18/09/2006, from http://plus.maths.org/issue14/features/smith.
JumpStart. (2010). "JumpStart." 2010, from http://jumpstart.ecs.soton.ac.uk/09-10/js.php?student=ug&action=home.
Kaplan, A (1998, 1964). Validity measurement. The Conduct of Inquiry: Methodology for Behavioural Science. A. Kaplan. Edison NJ, Transaction publishers**:** 198-199.
Karau, S J and K D Williams (1993). "Social loafing." Journal of Personality and Social Psychology 65(4): 681-706.
Kates, S M (2002). "Barriers to Deep Learning in Student Marketing Teams." Australasian Marketing Journal 10(2): 14-25.
Katzenbach, J R and D K Smith (1993a). "The Discipline of teams." Harvard Business Review 71(2): Ill-120.
Katzenbach, J R and D K Smith (1993b). The wisdom of teams. London, McGraw-Hill.
Kaufman, A S. (1994). "Practice Effects " Retrieved 17 March 2010, from http://www.speechandlanguage.com/cafe/13.asp.
Kaufman, D B and R M Felder (2000). "Accounting for individual effort in cooperative learning teams." Journal of Engineering Education 89(2): 133-140.
Keele. (2005). "Keele University Undergraduate Prospectus 2005, Modern Languages: French." Retrieved 15 July, 2004, from http://www.keele.ac.uk/undergraduate/prospectus/2005/textonly/dhcourses/french.htm
Kemler, B and P Thomson (2006). Helping Doctoral Students Write. London, Routledge.
Kench, P L, N Fielda, M Aguderaa and M Gilla (2009). "Peer assessment of individual contributions to a group project: Student perceptions." Radiography 15(2): 158-165.
Kerr, N L and S E Bruun (1983). "Dispensability of member effort and group motivational losses." Journal of Personality and Social Psychology 44(1): 78-94.
Kingston, P (2004). Fashion victims. The Guardian.
Kingston, P (2005). Make the most of what you've got. New Statesman, Special Suplement**:** xxii-xxiv.
Knight, J (2004). "Comparison of Student Perception and Performance in Individual and group Assessment in Practical Classes." Journal of Geography in Higher Education 28 (1): 63-81.
Knight, J (2008b). Re: 2004 Comparison of student perception study. R. Almond.
Knight, P, Ed. (1995). Assessment for learning in higher education. London, Kogan Page.
Knight, P and M Yorke (2008). "Assessment close up: The limits of exquisite descriptions of achievement." International Journal of Educational Research 47: 175-183.
Knight, P T (2000). "The Value of a Programme-wide Approach to Assessment." Assessment & Evaluation in Higher Education 25(3): 237-251.
Knight, P T (2002). "Summative Assessment in Higher Education: practice in disarray." Studies in Higher Education: 275 - 286.
Koretz, D, B Stecher, S Klein, D McCaffrey and E Deibert (1993). Can Portfolios Assess Student Performance and Influence Instruction? The 1991-92 Vermont Experience. CSE Technical Report 371, RAND Institute on Education and Training. National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
Kreft, I and J d Leeuw (1998). Introducing multilevel modeling. London, Sage.
Lambert, R (2003) Lambert Review of Business-University Collaboration
Latané, B, K Williams and S Harkins (1979). "Many Hands Make Light the Work: The Causes and Consequences of Social Loafing." Journal of Personality and Social Psychology 37(6): 822-832.
Laughlin, P R, E C Hatch, J S Silver and L Boh (2006). "Groups Perform Better Than the Best Individuals on Letters-to-Numbers Problems: Effect of Group Size." Journal of Personality and Social Psychology 90(4): 644-651.
Leathwood, C, L Johnson and S Moore (1999). Researching the roles of group work in learning, teaching and assessment: a comparative case study of two degree programmes. 7th international improving student learning symposium, University of York.
Leckey, J F and M A McGuigan (1997). "Right Track - Wrong Rails." Research in Higher Education 38(3): 365-378.
Lee, S (2004). Curriculum matters. LTSN conference on curriculum in the 21st century.
Lejk, M, M Wyvill and S Farrow (1996). "A Survey of Methods of Deriving Individual Grades from Group Assessments." Assessment & Evaluation in Higher Education 21(3): 267-280.
Lejk, M, M Wyvill and S Farrow (1997). "Group Learning and Group Assessment on Undergraduate

Computing Courses in Higher Education in the UK: results of a survey." Assessment & Evaluation in Higher Education 22(1): 81-91.
Lejk, M, M Wyvill and S Farrow (1999). "Group Assessment in Systems Analysis and Design: a comparison of the performance of streamed and mixed-ability groups." Assessment & Evaluation in Higher Education 24(1): 5-14.
LEMMA. (2009). "Learning Environment for Multilevel Methods and Applications." Retrieved 02 05 2009, from http://www.cmm.bris.ac.uk/lemma/.
Li, L K Y (2001). "Some Refinements on Peer Assessment of Group Projects." Assessment & Evaluation in Higher Education 26(1): 5-18.
Light, R and D B Pillemer (1984). Summing up. Cambridge, Harvard UP.
Linde, D v d (2002). A degree of doubt. The Guardian.
Lipsey, M (2003). "Those Confounded Moderators in Meta-Analysis." The Annals of the American Academy of Political and Social Science 587(69): 69-81.
Loddington, S (2008). Peer assessment of group work: a review of the literature, Loughborough University.
Lofland, J and L H Lofland (1995). Analysing social settings: a guide to qualitative observation and analysis. Belmont, CA, Wadsworth.
LondonMet, U. (2006). "University assessment framework." Retrieved 10 07 2006, from http://www.londonmet.ac.uk/library/c71911_3.pdf.
Luft, J (1984). Group Processes: A Introduction to Group Dynamics (3rd Edition). Mountain view CA, Mayfield.
MacBean, J, T Graham and C Sangwin. (2004). "Guidelines for introducing group work in undergraduate mathematics." Retrieved 21 April, 2004, from http://ltsn.mathstore.ac.uk/projects/groupwork/intro.htm
Mahoney, K (2008). "Linguistic Influences on Differential Item Functioning for Second Language Learners on the National Assessment of Educational Progress " International Journal of Testing 8(1): 14-33.
Maiden, B and B Perry. (2010). "Dealing with free-riders in assessed group work: results from a study at a UK university." Assessment & Evaluation in Higher. First published on: 04 February 2010 (iFirst), from http://dx.doi.org/10.1080/02602930903429302.
Maranto, R and A Gresham (1998). "Using "World Series shares" to fight free riding in group projects.(teaching of political science)." Political Science & Politics 31(4): 789(3).
May, T (2001). Social Research: issues, methods and process (3rd Ed.). Buckingham Open University Press.
Mayer, E (1992). Report of the committee to advise the Australian education council and ministers of vocational education, employment and training on employment-related key competencies for postcompulsory education and training.
McKenna, G (1995). "Learning theories made easy: humanism." Nursing Standard 9(31): 29-31.
Meadows, M and L Billington (2005). A Review of the Literature on Marking Reliability, National Assessment Agency.
Mello, J A (1993). "Improving individual member accountability in small work group settings." Journal of Management Education 17(2): 253-259.
Messick, S (1980). "Test validity and ethics of assessment." American Psychologist 35(11): 1012-1027.
Messick, S (1993). Validity. Educational measurement. R. L. Linn. Phoenix, Oryx press**:** 13-103.
Michaelson, R. (2004). "Assessing Group Work." Retrieved March, 2004, from http://cbs1.gcal.ac.uk/lts/Word%20files/Briefing%20Papers/RMGroupwork%20web%20-finalR.doc
Miller, A H, B W Imrie and K Cox (1998). Student assessment in higher education. London, Kogan Page.
Moore, E C (1915). What is education. Boston, Ginn and Company.
Moreland, N (2004). Learning & Employability. Employability: Work-related learning in higher education. York, The Higher Education Academy (LTSN).
Morgan, P. (2005a). "Improving the Group Working Experience: The Effectiveness of Group Work Training." Retrieved 20/01/2005, from www.herts.ac.uk/envstrat/HILP/conferences/3rd/PAPERS/Morgan.doc.
Morris, R and C Hayes (1997). Small Group Work: Are group assignments a legitimate form of assessment? 6th Annual Teaching Learning Forum, Perth, Murdoch University.
Murray, M H. (2003). "Managing teamwork on-line." Retrieved 23 August 2004, from http://eprints.qut.edu.au/archive/00000080/01/MartinMurray.PDF
Mutch, A (1998). "Employability or learning? Groupwork in higher education." Education & Training 40(2): 50-56.
Mutch, A and G Brown (2001). Assessment: A guide for heads of departments, LTSN Generic Centre.
Myers, R (2007). Team Project module interview. R. Almond.

Myron-Wilson, R and P K Smith (1998). "A matter of degrees." The Psychologist: 535-538.
Newstead, S (2002). "Examining the examiners: Why are we so bad at assessing students?" Psychology Learning and Teaching 2(2): 70-75.
Newstead, S (2008). Slide Effect. R. J. Almond.
Newton, P E (2007). "Clarifying the purpose of educational assessment." Assessment in Education 14(2): 149-170.
Nicholl, P and S Alexander (2004a). A Study Into Managing Group Work When Dealing With Large Student Numbers, Higher Education Academy**:** 156-162.
Nicolay, J A (2002). "Group Assessment in the On-Line Learning Environment." New Directions For Teaching and Learning 91: 43-52.
Noble, A, C Ingleton, L Doube and T Rogers. (undated). "Leap into Collaborative Learning." Retrieved 3rd April 2002, from http:/www.adalaide.edu.au/ltdu/leap/leapinto/collab_learning/index.html
Novick, M R (1985). Standards for educational and psychological testing. Washington DC, APA.
Oakley, B. (2004). "Problem mini-clinics - Empowering students to deal with troublesome teammates." Retrieved 14 11 2006, from http://www.oncourseworkshop.com/On%20Course%20Newsletter.htm.
Oldfield, S J. (2005). "Making a Virtue out of Virtuality: Team Working in Distributed Environments." Retrieved December 2008, from http://ecet.ecs.ru.acad.bg/cst05/Docs/cp/sIV/IV.1.pdf.
Olson, M (1965). Public Goods and the Theory of Groups. Cambridge MA, Harvard University Press.
Oppenheim, A N, M Jahoda and R L James (1967). "Assumptions underlying the use of university examinations." Universities Quarterly 21(3): 341-351.
Orr, S (2010). "Collaborating or fighting for the marks? Students' experiences of group work assessment in the creative arts." Assessment & Evaluation in Higher Education 35(3): 301-313.
Palmer, J A (2001). Fifty major thinkers on education from Confucius to Dewey. London, Routledge.
Paton, G (2009). Traditional degree grades should be scrapped as part of an overhaul of university standards, according to an official report. . Daily Telegraph.
Paulsen, F (1908). The German Universities and university study, University of Toronto.
Paxton, T (1963). I Can't Help But Wonder Where I'm Bound, Elektra. EKL-277.
Penny, A R and R Coe (2004). "Effectiveness of Consultation on Student Ratings Feedback: A Meta-Analysis " Review of Educational Research 74(2): 215-253.
Peugh, J L (2010). "A practical guide to multilevel modeling (sic)." Journal of School Psychology 48: 85-112.
Phillips, D C and N C Burbules (2000). Postpositivism and educational research. Lanham, Maryland, Rowman and Littlefield.
Pitt, M J (2000). "The Application of Games Theory to Group Project Assessment." Teaching in Higher Education 5(2): 233-241.
Race, P (1995a). "The Art of Assessment " The New Academic 4(3).
Race, P (1998). 500 Tips for Open and Flexible Learning. London, Kogan Page.
Ramsden, P (1992). Learning to teach in higher education. London, Routledge Falmer.
Ramsden, P (2003). Learning to teach in higher education (2nd ed.). London, Routledge Falmer.
Rasbash, J. (2008). "Module 4: Multilevel structures and classifications." 2009, from www.cmm.bristol.ac.uk.
Rasbash, J, F Steele, W Browne and H Goldstein (2009). A User's Guide to MLwiN Version 2.10. Bristol, Centre for Multilevel Modelling, University of Bristol.
Richardson, M. (2004). "Financial Mail Lunch Speech." Retrieved 21 August 2007, from http://www.hsbc.com/1/PA_1_1_S5/content/assets/csr/speech_dame_mary_richardson_2004.pdf.
Ritter, L (2000). "The Quest for an Effective Form of Assessment: the evolution and evaluation of a controlled assessment procedure (CAP)." Assessment & Evaluation in Higher Education 25(4): 307-320.
Rogers, C R (1983). Freedom to learn for the 80's. Columbus OH, Merrill.
Rogers, J (2001). Adults Learning (4th Ed.). Buckingham, OU.
Rothwell, B A. (2002). "Peer Evaluations, and Team Learning In Undergraduate and Graduate Education." Retrieved 30 March, 2004, from http://www.ec.erau.edu/cce/facsen0202/paper4.doc
Rowntree, D (1987). Assessing Students: how shall we know them? second edition. London, Kogan Page.
RSA (1995). Tomorrow's Company, The Royal Society for the Encouragement of Arts, Manufacturers and Commerce.
Ruch, G M (1933). "Recent Developments in Statistical Procedures." Review of Educational Research 3(1): 33-40.

Salomon, G (1992). "What does the design of effective CSCL require and how do we study its effects?" ACM SIGCUE Outlook 21(3): 62-68.
Sambell, K, L McDowell and S Brown (1997). ""But is it fair?" : An exploratory study of student perceptions of the consequential validity of assessment." Studies in Educational Evaluation 23(4): 349-371.
Sanders, M. (2008). "A Failure to Collaborate." Retrieved 20/02/2008, from http://chronicle.com/jobs/news/2008.02/2008021801c/printable/html.
Schullery, N M and S E Schullery (2006). "Are heterogeneous groups more beneficial to students:." Journal of Management Education 30(4): 542-556.
Scriven, M (1973). The Methodology of Evaluation. Educational Evaluation: Theory and Practice. B. R. Worthen and J. R. Sanders. Worthington Ohio, Charles A Jones**:** 60-106.
Scriven, M (1983). Evaluation Ideologies. Models and Conceptualizations. G. F. Madaus, M. Scriven and D. L. Shufflebeam. Boston, Kluwer-Nijhoff Publishing**:** 229-260.
Scriven, M (2004). Reflections. Evaluation Roots: tracing theorists' views and influences. M. C. Alkin. Thousand Oaks, Sage**:** 183-195.
Sharp, S (2006). "Deriving individual student marks from a tutor's assessment of group work." Assessment and Evaluation in Higher Education 31(3): 329-343.
Sheard, A G and A P Kakabadse (2004). "A process perspective on leadership and team development." Journal of Management development 23(1): 7-106.
Shields, P M and H Tajalli (2006). "Intermediate Theory: The Missing Link to Successful Student Scholarship." Journal of Public Affairs Education, Faculty Publications-Political Science. Paper 39. http://ecommons.txstate.edu/polsfacp/39 12(3): 313-334.
Simonite, V (2000). "The Effects of Aggregation Method and Variations in the Performance of Individual Students on Degree Classifications in Modular Degree Courses." Studies in Higher Education 25(2): 197-209.
Simonite, V (2001). An application of multilevel modelling techniques to the longitudinal study of student progress in a modular degree course. Institute of Education, University of London. PhD**:** 279 pages.
Simonite, V (2002). "Finding the weak spots: Should some students' programmes carry a health warning?" Teaching forum 50(Autumn 2002): 48-49.
Simonite, V (2003a). "The Impact of Coursework on Degree Classifications and the Performance of Individual Students." Assessment & Evaluation in Higher Education 28(5): 459-470.
Simonite, V (2003b). "A Longitudinal Study of Achievement in a Modular First Degree Course." Studies in Higher Education 28(3): 293-302.
Simonite, V and W J Browne (2003c). "Estimation of a large cross-section multilevel model to study achievement in a modular degree course." Journal of the Royal Statistics Society 166(Part 1): 119-133.
Sluijsmans, D and F Prins (2006). "A Conceptual Framework for Integrating Peer Assessment in Teacher Education." Studies in Educational Evaluation 32: 6-22.
Smith, A (1976). The Wealth of Nations. Glasgow.
Smith, M K. (2008a). "Eduard C Lindeman and the meaning of adult education." Retrieved 31 12 2008, from http://www.infed.org/thinkers/et-lind.htm.
Smith, M K. (2008b). "Andragogy." Retrieved 31 12 2008, from http://www.infed.org/lifelonglearning/b-andra.htm.
Smithers, R. (2006). "Exam regulator publishes report to stop coursework cheats." Retrieved 29 March 2009, from http://www.guardian.co.uk/education/2006/mar/03/schools.uk1.
Snyder, B S (1971). The hidden curriculum. New York, Knopf.
Stobart, G (2008). Testing times: the uses and abuses of assessment. London, Routledge.
Strauss, A and J Corbin (1990). Basics of Qualitative Research: grounded theory procedures and techniques. London, Sage.
Strauss, P (2001). ""I'd rather vomit up a live hedgehog" - L2 students and group assessment in mainstream university programmes." Australian Journal of TESOL 16(2).
Sydney, U o T. (1999). "Student Groups: Issues for Teaching and Learning." Retrieved 18 June, 2004, from http://www.clt.uts.edu.au/Student.Groupwork.html
Tait, K (2009). "Reflecting on How to Optimize Tertiary Student Learning Through the Use of Work Based Learning Within Inclusive Education Courses." International Journal of Teaching and Learning in Higher Education 20(2): 192-197.
Taras, M (2005). "Assessment - Summative and Formative - Some Theoretical Reflections." British Journal of Educational Studies 53(4): 466-478.
Taylor, C (1985). Philosophical Papers, Cambridge University Press.
Taylor, T and A Rumpus (1997). Using Alumni to Formulate Staff Development for the Design and Delivery of a Curriculum Relevant to employment, University of Westminster**:** 31.
Terms. (2009). "Keeping of Terms." Retrieved 08 04 09, from http://www.dur.ac.uk/resources/university.calendar/volumei/2005.2006/regulations/reg5.pdf .

Thompson, D and I McGregor (2005). Self and Peer Assessment for Group Work in Large Classes. Making a Difference: 2005 Evaluations and Assessment Conference, Sydney.
Thorley, L and R Gregory (1994). Using Group-based Learning in Higher Education. London, Kogan Page.
Townend, S. (1997). "Group Work and Peer Tutoring." Retrieved 30 March, 2004, from http://www.hull.ac.uk/mathskills/newsletters/issue4/page8.htm
Trafford, V and S Leshem (2002). "Starting at the end to undertake doctoral research: Predictable questions as stepping stones." Higher Education Review 34(1): 31-49.
TRAMSS. (1999). "MLwiN - What is Multilevel Modelling?" Retrieved 25 January 2010, from http://tramss.data-archive.ac.uk/documentation/MLwiN/what-is.asp.
Tuckman, B W (1965). "Developmental sequence in small groups." Psychological Bulletin 63(6): 384-99.
Tymms, P (1995). "The long-term impact of schooling." Evaluation & Research in Education 9(2): 99-108.
UCL. (2004). "Working effectively in a group (BiolB235)." Retrieved 15 July, 2004, from http://www.ucl.ac.uk/keyskills/customised-pages/biology/maptextonly.html
Ulster, U. (2004). "The Peer Learning in Music Project." Retrieved 25 05 2004, from http://www.ulst.ac.uk/faculty/humanities/mpa/html/plm.intro.htm#anchor130399
University of Central Lancashire. (2004). "Extraterrestrial life." Retrieved 01/06/2004, from http://www.uclan.ac.uk
University of Western Cape, U. (2004). "Studying Zoology at UWC." Retrieved 15 July 2004, from http://www.science.uwc.ac.za/zoology/study.htm
UUK (2010). Higher Education in Facts and Figures (Summer 2010), Universities UK.
Watson, S B and J E Marshall. (1995b). "Heterogenous grouping as an element of cooperative learning in an elementary education science course." Retrieved 23 October 2009.
Webb, N. (1994). "Group Collaboration in Assessment: Competing objectives, processes and outcomes." Retrieved 9 10 2006, from http://www.cse.ucla.edu/Summary/386webb.htm.
Wellington, V U o. (2004). "Improving teaching and Learning: Group Work and Group Assessment." Retrieved 06/09/2006, from http://www.vtde.vuw.ac.nz/resources/guidelines/Groupwork.pdf.
Welton, J (1914). What do we mean by education? London, Macmillan and Co.
White, S (2006). Using groupwork to enhance first experiences at university. HEA Group Work Workshop. Computer Science Department, University of Durham.
White, S and L Carr (2005c). Brave New World: Can We Engineer a Better Start for Freshers? 35th ASEE/IEEE Frontiers in Education Conference, Indianapolis, IEEE.
Wiliam, D and P Black (1996). "Meanings and Consequences: A Basis for Distinguishing Formative and Summative Functions of Assessment?" British Educational Research Journal 22(5): 537-548.
Wilson, N. (1998). "Educational Standards and the Problem of Error." Retrieved 21 Jan., 2008, from http://epaa.asu.edu/epaa/v6n10/c1.htm.
Wolf, A (1991). "Assessing Core Skills: wisdom or wild goose chase?" Cambridge Journal of Education 21(2): 189-201.
Wolf, A (2002). Does education matter? Myths about education and economic growth. London, Penguin.
Wollongong, U o. (2006). "Learning and Teaching\ Good Practice: Assessment\ B%. Group Work." Retrieved 26 01 2006, from http://www.uow.edu.au/aboput/teaching/goodpractice/assessment/B5groupwork.html
Yero, J L. (2002). "The meaning of education." Retrieved 15 06 2007, from http://www.TeachersMind.com/pdfdirectory/Education.PDF.
York, U o. (2002). "Introduction to world music." Retrieved 15 July, 2004, from http://www.york.ac.uk/admin/sro/modules/ugrad/music.htm
Yorke, M (2007). The law of averages produces poor results. THES.
Young, C B and J A Henquinet (2000). "A conceptual Framework for Designing Group Projects." Journal of Education for Business: 56-60.
Zwanenberg, N v, L J Wilkinson and A Anderson (2000). "Felder and Silverman's Index of Learning Styles and Honey and Mumford's Learning Styles Questionnaire: how do they compare and do they predict academic performance?" Educational Psychology 20(3): 365-378.

Appendixes

Appendix 1. Principal stakeholders of GSA

Parents
Sixth form and other Further Education students
Potential undergraduate students
Undergraduate and postgraduate students
School staff, including teachers, careers guidance staff and governors
FE institutions
Local Education Authorities
Academics, university teaching staff, and module convenors and administrators
Graduate recruiters and employers
Departments of local and national government
Quality agencies
Funding bodies
Compilers of league tables
Taxpayers

Appendix 2. Reasons for HE study

Peer pressure
School pressure/tradition
Family pressure/tradition
To confound peer, school and/ or family pressure
Professional/domestic opportunity
To delay alternative actions
Because it is a first generation opportunity
To facilitate entry to a professional career
For the social opportunities
For the sporting opportunities
For the networking opportunities
For the travel opportunities
As a personal challenge

Appendix 3. Attributes of a Durham honours degree graduate

HONOURS DEGREE
1. A Durham Honours graduate will have developed an understanding of a complex body of knowledge, some of it at or close to current boundaries of an academic discipline or disciplines. The graduate will be able to evaluate evidence, arguments and assumptions, to reach sound judgements, and to communicate effectively.

2. A Durham Honours degree is awarded to a student who has demonstrated:
(a) a systematic understanding of their field of study, based upon coherent knowledge of core areas, and advanced knowledge of selected aspects that is informed by research at, or close to, the current forefront of the subject(s);
(b) The ability to deploy accurately established techniques of analysis and enquiry within the subject(s);
(c) Conceptual understanding that enables the student:

(i) To devise and sustain arguments, and/or to solve problems, using ideas and techniques, some of which are at the forefront of a discipline;
(ii) To describe and comment upon particular aspects of current research, or equivalent advanced scholarship, in defined areas of their programme of study;
(iii) To use appropriate theoretical and conceptual frameworks to order and/or interpret new data or kinds of evidence;
(iv) To critically evaluate the reliability, validity and significance of data, evidence or interpretations within the field of study;

(d) An awareness of current disciplinary boundaries and an appreciation of the uncertainty limits and contested nature of knowledge within their programme of study;
(e) The ability to direct and manage their own learning effectively across a range of topics;
(f) The ability to make use of scholarly reviews and primary sources (e.g. refereed research articles and/or original materials appropriate to the discipline);
(g) The ability to undertake, with supervision, independent investigation of a defined topic within their programme of study and to report the findings effectively.

3. Typically, a holder of the qualification will be able to:
(a) Apply the methods and techniques that they have learned to review, consolidate, extend and apply their knowledge and understanding; and to initiate and carry out projects or further enquiry within areas of their programme of study;
(b) Critically evaluate arguments, assumptions, abstract concepts and data (that may be incomplete); to formulate judgements, and to frame appropriate strategies to investigate topics within defined aspects of their programme of study;
(c) Communicate relevant information, concepts, ideas and, where appropriate, problems and solutions, within their programme of study to both specialist and non-specialist audiences; and will have:
(d) Academic qualities such as flexibility and discrimination, together with personal qualities and transferable skills necessary for employment requiring:
(i) The exercise of initiative and personal responsibility;
(ii) Decision making in complex and unpredictable contexts;
(iii) The learning ability needed to undertake appropriate further education and training of a professional or equivalent nature and standard.

At the foot of the page was the note that:

> *"These qualification descriptors are generic and apply to all subject areas at the relevant level across the University. Each department supplements these with its own subject-specific descriptors in line with the relevant benchmarks and other requirements appropriate to the discipline"*

(Descriptors 2008)

Appendix 4. Slide-effect

> *"Farrell and Gilbert (1960) argued that the variance of the marks an examiner awards will increase relative to the number of scripts he or she has already marked because of either growth in confidence or examiner fatigue. They tested the hypothesis that the more scripts an examiner marks the more likely he or she will be to award extreme marks. The undergraduate scripts were marked in alphabetical order, so it was predicted that extreme marks would occur most frequently in the later part of the alphabet. Unfortunately, Farrell and Gilbert only had access to the classification awarded to the scripts rather than the mark. They categorised the classifications as being either central (upper second, lower second and third class) or extreme (first class or below third class). Each candidate being classified according to whether his or her grade was extreme or central, and whether the initial of his or her surname came before L or after K in the alphabet. A small but highly significant effect of the sort predicted was found."*

(Meadows and Billington 2005:21).

Appendix 5. Bibliography of methods of deriving individual marks

Allen, J. and Lloyd-Jones, R. (2006)
Barton et al. (2002)
Black, P. 1998 (1998)
Boud, D., Cohen, R. and Sampson, J. (2001)
Brown, R. W. (1995)
Brown, S. and Knight, P. (1994)
Burd, E., Drummond, S. and Hodgson, B. (2003)
Conway et al. (1993)
Earl, S. E. (1986)
Falchikov, N. (1986)
Fellenz, M. R. (2006)
Gibbs, Habashaw and Habashaw (1988)
Goldfinch, J. (1994)
Goldfinch, J. and Raeside, R. (1990)
Grajczonek, J. (2009)
Greenan, K. and Alexander, S. (2003)
Heathfield, M. (1999)
Jaques, D. (2000)
Knight, J. (2004)
Lejk, M., Wyvill, M. and Farrow, S. (1996)
Lejk, M., Wyvill, M. and Farrow, S. (1997)
Li, L. K. Y. (2001)
Noble et al. (undated)
Rothwell, B. A. (2002)
Sharp, S. (2006)
Thompson, D. and McGregor, I. (2005)
Thorley, L. and Gregory, R. (1994)
Wellington, V. U. o. 2004 (2004)

Appendix 6. Reasons for group project work

Reasons for group project work

To increase the amount and quality of discussion between students and foster informal peer tutoring and peer feedback

To enable students to be involved in larger scale, more complex and more open-ended learning tasks than they could manage on their own

To produce better quality learning outcomes than any individual student could manage

To increase cooperation between students within groups, but with competition between groups, in order to produce maximum overall productivity

To develop students' teamwork skills but also to involve a whole range of other transferable skills, such as time and task management, creative problem solving and written and oral communication skills

To save resources: Group work can be more economical to set up, supervise, equip and mark than individually undertaken project work

(Gibbs 1995a)

Appendix 7. Purposes of assessment(Much and Brown 2001:5)

Learning	To provide feedback to students to improve their learning
	To motivate students
	To diagnose a student's strengths and weaknesses
	To help students to develop their skills of self-assessment
	To provide a profile of what a student has learnt
Certification	To pass or fail a student
	To grade or rank a student
	To licence to proceed
	To licence to practice
	To select for future courses
	To predict success in future courses
	To select for future employment
	To predict success in employment
Quality Assurance	To provide feedback to lecturers on student learning
	To improve teaching
	To evaluate a course's strengths and weaknesses
	To assess the extent to which a programme has achieved its aims
	To judge the effectiveness of the learning environment
	To ensure the course is credit worthy to other institutions
	and employers
	To monitor standards over time

Appendix 8. Atkins et al. six flaws in assessment practice, cited by Elton (2004)

> No consistency in criteria used between subjects, within subjects, between institutions and within institutions for awarding of degree class.
>
> Certain frames of reference that lecturers bring to assessment are systematically biased, but the bias is often subconscious and unrecognised.
>
> Internally, lecturers have little idea of how others set and mark assignments; external examiners are not usually part of the curriculum design team; both are usually untrained in assessment.
>
> Few lecturers understand the technical design factors that can affect assessment outcomes.
>
> New forms of assessment, e.g. continuous assessment, are as prone to distortion as formal examinations.
>
> Although there are exceptions, in many departments the approach to assessment remains conservative through ignorance or unwillingness to consider change.

(Elton 2004:43-44)

Appendix 9. Synthesized data set B6

Data set B5, synthesized from figure 7 in Knight (2004:74), see for example 5.1.2.

The student ranked 37 had the same score for both categories of marks (62). This is the x-axis bisection of the regression lines.

IISA Mark Rank	IISA Mark	GSA Mark	IISA Mark Rank	IISA Mark	GSA Mark
1	14	67	28	57	57
2	18	70	29	59	70
3	28	62	30	60	71
4	31	49	31	60	70
5	33	65	32	60	50
6	34	60	33	62	68
7	35	68	34	62	67
8	39	35	35	62	65
9	40	54	36	62	63
10	40	68	37	62	62
11	40	69	38	65	68
12	40	71	39	65	52
13	45	58	40	66	52
14	47	62	41	68	52
15	47	38	42	68	71
16	48	63	43	69	54
17	49	45	44	69	47
18	50	51	45	69	72
19	50	65	46	71	73
20	51	51	47	74	78
21	51	63	48	77	82
22	53	57	49	77	80
23	54	62	50	77	60
24	55	82	51	78	64
25	56	77	52	79	80
26	56	52	53	85	72
27	57	63			

Appendix 10.C data group allocation algorithm example

This example was taken from Almond (2006). There were 60 students in the cohort so 10 groups of group 6 members were required. Students were first listed 1-60 by ability from previous course scores.

From the top down the first half of the list 1-30 was numbered sequentially 1, 2, 3, 4, 5, 6, 1, 2, 3, etc. i.e. the group they are allocated to. Then, starting at the end of the list (student 60), the remainder were given similar numbering.

This method of allocating students to mixed ability groups will be fine where that ability difference is approximately linear. In practice, such a variable attribute is immeasurable.

Problems could occur if, for example, the externally based Natural Science students could not attend the teaching sessions because of conflicting class schedules, so they would have to move to a more conveniently timetabled group. Problems could also occur because of student dropout, which, anecdotally, inevitably seems to happen in the smallest groups. The aim was for groups of six. Sometimes the numbers were only four.

C Data Set example of group allocation algorithm

StuID by previous ability score	allocated to group		StuID by previous ability score	allocated to group		StuID	allocated to group
Stu1	1		Stu31	10		Stu1	1
Stu2	2		Stu32	9		Stu11	1
Stu3	3		Stu33	8		Stu21	1
Stu4	4		Stu34	7		Stu40	1
Stu5	5		Stu35	6		Stu50	1
Stu6	6		Stu35	5		Stu60	1
Stu7	7		Stu37	4		…	
Stu8	8		Stu38	3		…	
Stu9	9		Stu39	2		…	
Stu10	10		Stu40	1			
Stu11	1		Stu41	10			
Stu12	2		Stu42	9			
Stu13	3		Stu43	8		Stu10	10
Stu14	4		Stu44	7		Stu20	10
Stu15	5		Stu45	6		Stu30	10
Stu16	6		Stu46	5		Stu31	10
Stu17	7		Stu47	4		Stu41	10
Stu18	8		Stu48	3		Stu51	10
Stu19	9		Stu49	2			
Stu20	10		Stu50	1			
Stu21	1		Stu51	10			
Stu22	2		Stu52	9			
Stu23	3		Stu53	8			
Stu24	4		Stu54	7			
Stu25	5		Stu55	6			
Stu26	6		Stu56	5			
Stu27	7		Stu57	4			
Stu28	8		Stu58	3			
Stu29	9		Stu59	2			
Stu30	10		Stu60	1			

Appendix 11. Convenor interview schedule

Course Code:		
Course Name:		
Contact Name:		
Dept/School:		
Faculty:		

This checklist is for background to help me to know the data, not part of the data.
Explain the study & MA pilot. [The institution central records department] will supply subject marks data.

Course components:
From On-Line Faculty Handbook: (Anonymised) summative assessment has n components.
Abc, nn%
Def, nn%
Ghi, nn% etc.
Is this how it works in practice?

How many students have there been over the last 3 years?

How many students are in each project group?

How is group membership determined? Self; Tutor; Random; Algorithm; Hetero- or Homogeneous Ability; Gender; Ethnicity

Is there 1 project topic; do students select from a list?

Has the course format, especially the group assessment weighting, remained stable, over the last 3 years?

How are individual marks derived from group marks? Tutor/Self/Peer/Same; algorithm; weighting?

Which programmes or departments use this module.

What is the rationale for using group rather than individual summative assessment?

What's your view on 'degrees are awarded to individuals, not groups'? e.g. not applicable; no view

What is your view on groups vs. teams? e.g. semantics; not applicable; no view …

How is group-dynamics taught? (e.g. not taught, Tuckman, Bion, Janis, Belbin, Myers-Briggs)

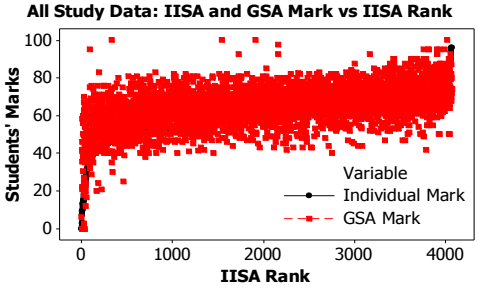
Have the group and individual course marks been compared elsewhere? Was it published? Where?

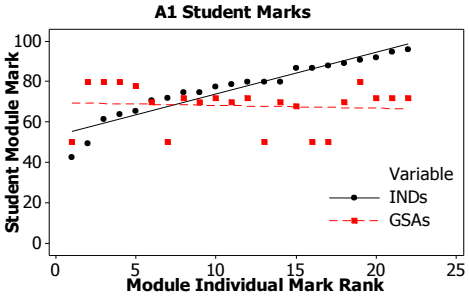
What are your thoughts on my including these course marks in my study?

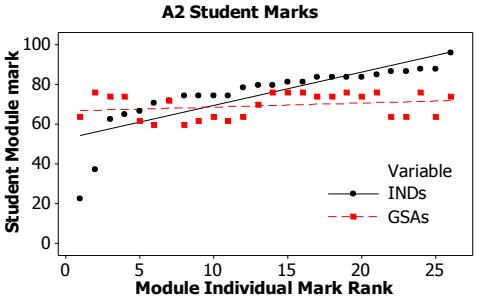
Which other group assessment courses do you know of? (I might have missed some!)

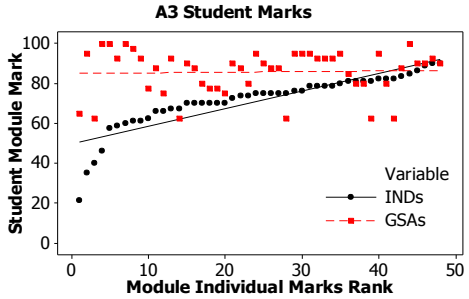
May I e-mail you later, with other queries, if necessary?

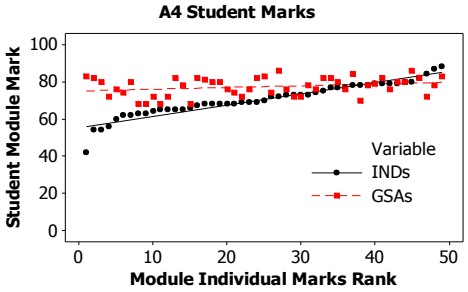
Nineteen Scatter Plots Including Data Points

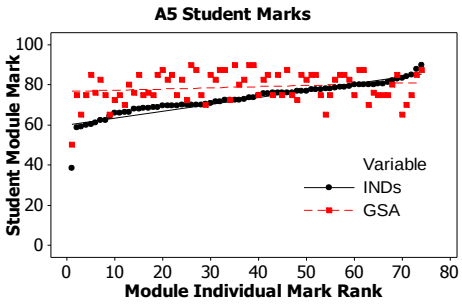


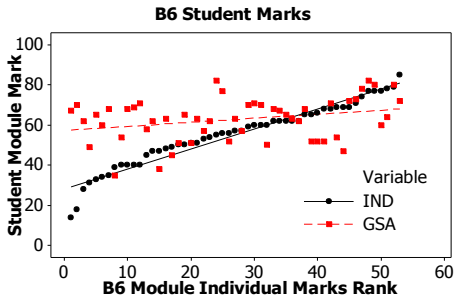


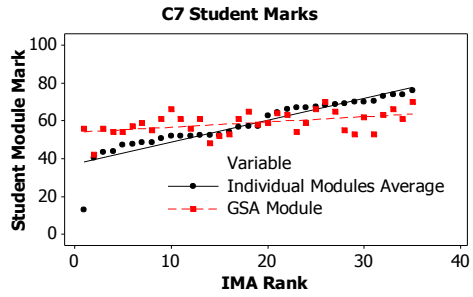


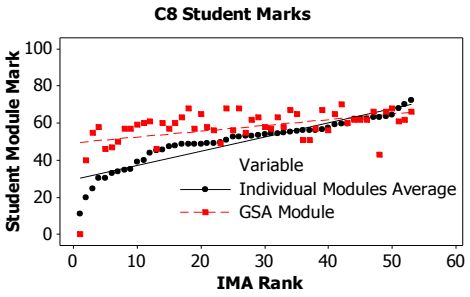


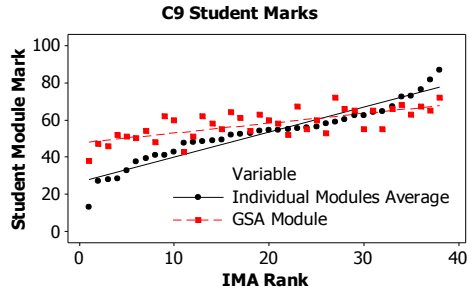


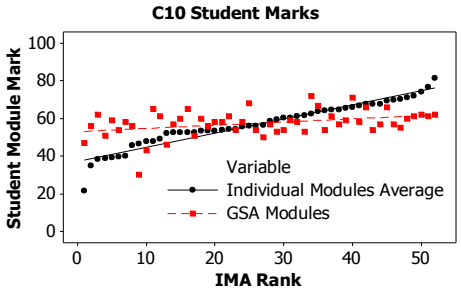


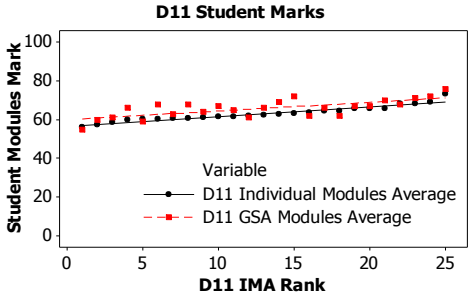


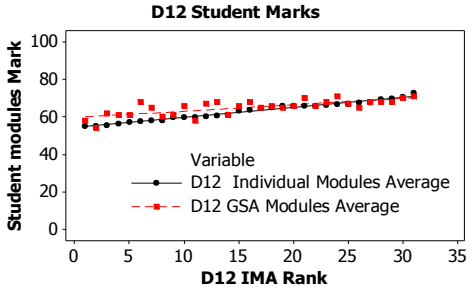

C7 Student Marks
Student Module Mark
IMA Rank
Variable
Individual Modules Average
GSA Module

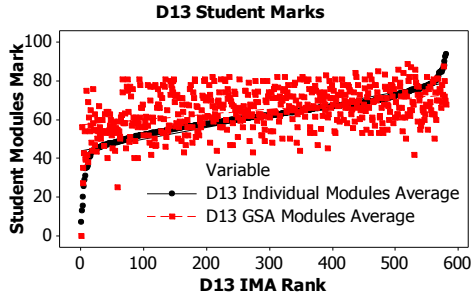

C8 Student Marks
Student Module Mark
IMA Rank
Variable
Individual Modules Average
GSA Module

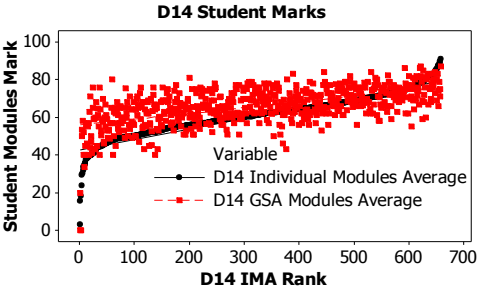

C9 Student Marks
Student Module Mark
IMA Rank
Variable
Individual Modules Average
GSA Module

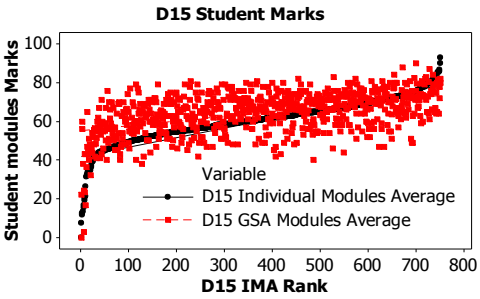

C10 Student Marks
Student Module Mark
IMA Rank
Variable
Individual Modules Average
GSA Modules

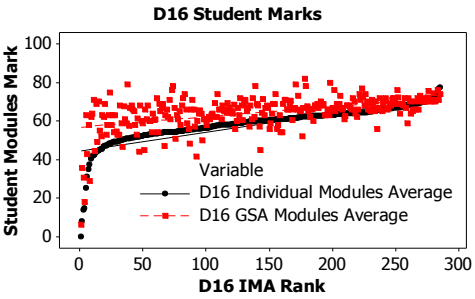

D11 Student Marks
Student Modules Mark
D11 IMA Rank
Variable
D11 Individual Modules Average
D11 GSA Modules Average

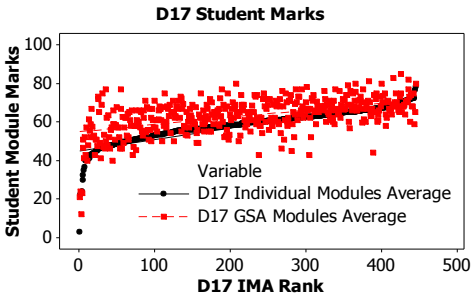

D12 Student Marks
Student modules Mark
D12 IMA Rank
Variable
D12 Individual Modules Average
D12 GSA Modules Average

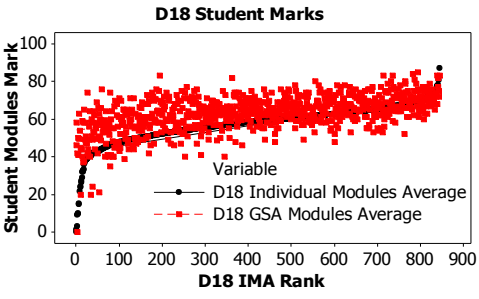

D13 Student Marks
Student Modules Mark
D13 IMA Rank
Variable
D13 Individual Modules Average
D13 GSA Modules Average


D14 Student Marks
Student Modules Mark
D14 IMA Rank
Variable
D14 Individual Modules Average
D14 GSA Modules Average

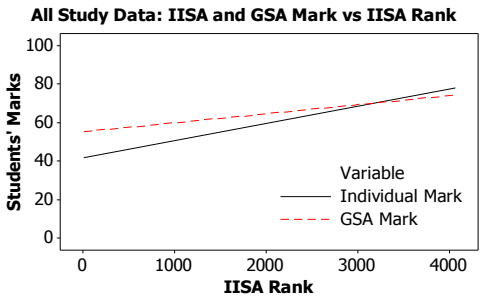

D15 Student Marks
Student modules Marks
D15 IMA Rank
Variable
D15 Individual Modules Average
D15 GSA Modules Average

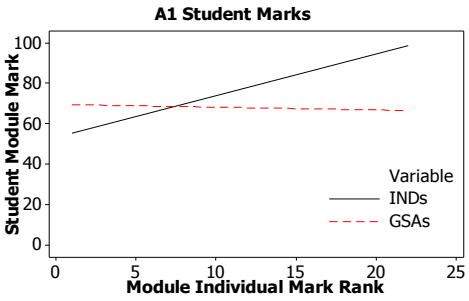

D16 Student Marks
Student Modules Mark
D16 IMA Rank
Variable
D16 Individual Modules Average
D16 GSA Modules Average

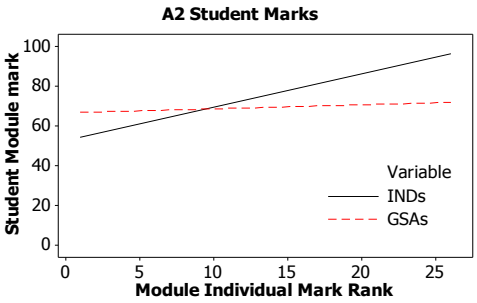

D17 Student Marks
Student Module Marks
D17 IMA Rank
Variable
D17 Individual Modules Average
D17 GSA Modules Average

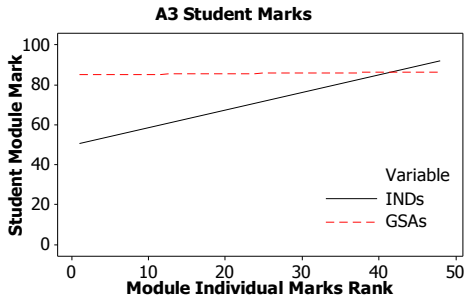

D18 Student Marks
Student Modules Mark
D18 IMA Rank
Variable
D18 Individual Modules Average
D18 GSA Modules Average

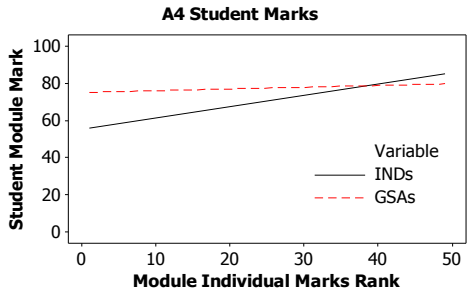
Nineteen scatter plots without data points

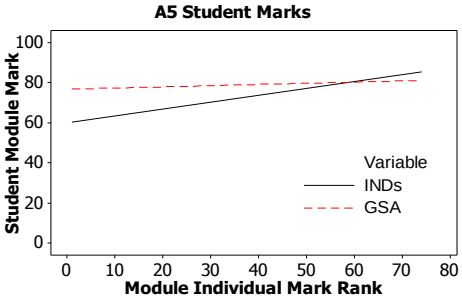


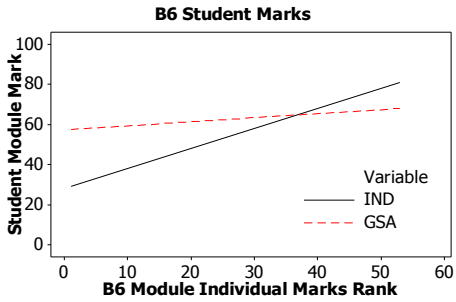


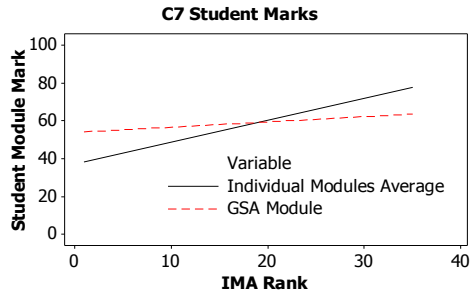


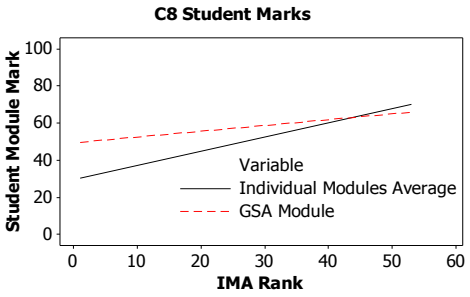


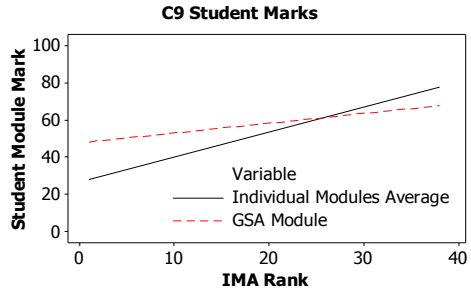


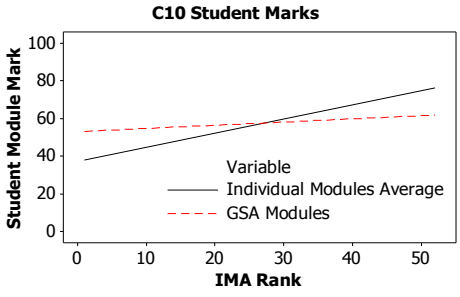


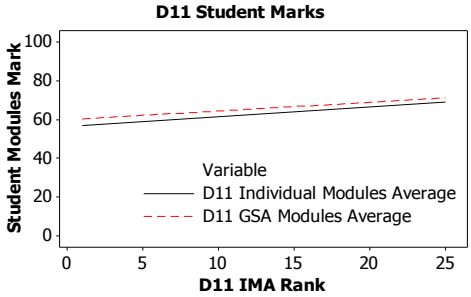


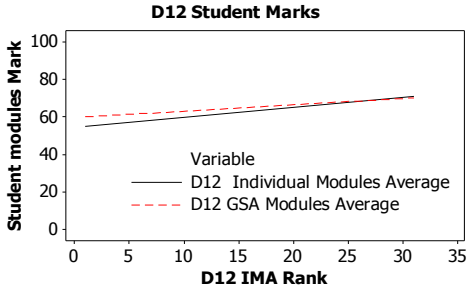

C7 Student Marks
Student Module Mark
IMA Rank
Variable
Individual Modules Average
GSA Module

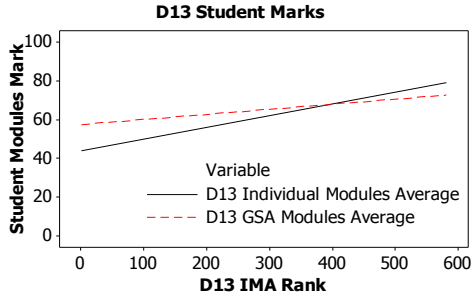

C8 Student Marks
Student Module Mark
IMA Rank
Variable
Individual Modules Average
GSA Module

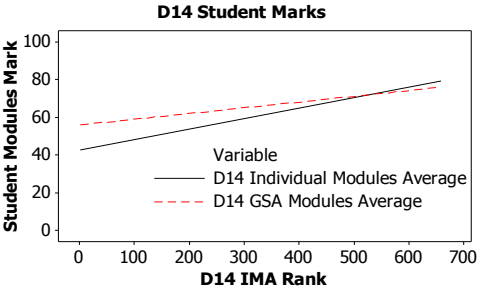

C9 Student Marks
Student Module Mark
IMA Rank
Variable
Individual Modules Average
GSA Module

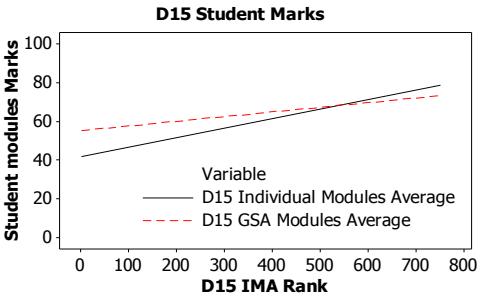

C10 Student Marks
Student Module Mark
IMA Rank
Variable
Individual Modules Average
GSA Modules

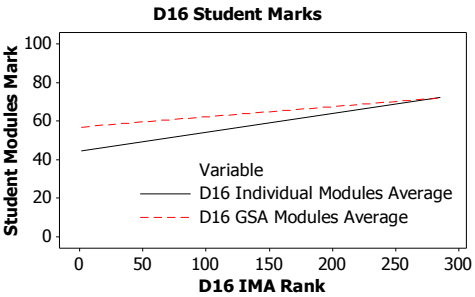

D11 Student Marks
Student Modules Mark
D11 IMA Rank
Variable
D11 Individual Modules Average
D11 GSA Modules Average

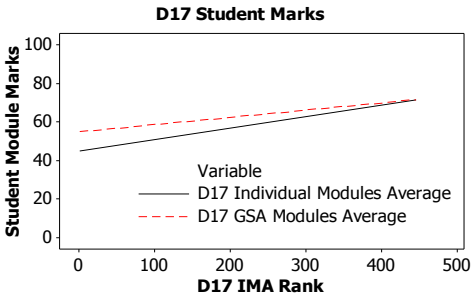

D12 Student Marks
Student modules Mark
D12 IMA Rank
Variable
D12 Individual Modules Average
D12 GSA Modules Average

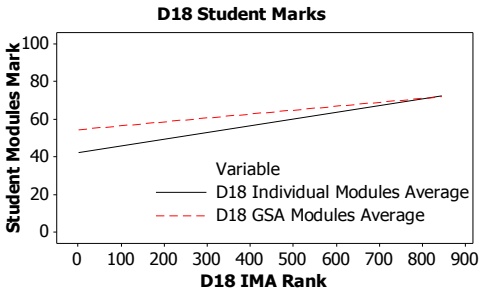

D13 Student Marks
Student Modules Mark
D13 IMA Rank
Variable
D13 Individual Modules Average
D13 GSA Modules Average


D14 Student Marks
Student Modules Mark
D14 IMA Rank
Variable
D14 Individual Modules Average
D14 GSA Modules Average


D15 Student Marks
Student modules Marks
D15 IMA Rank
Variable
D15 Individual Modules Average
D15 GSA Modules Average


D16 Student Marks
Student Modules Mark
D16 IMA Rank
Variable
D16 Individual Modules Average
D16 GSA Modules Average


D17 Student Marks
Student Module Marks
D17 IMA Rank
Variable
D17 Individual Modules Average
D17 GSA Modules Average


D18 Student Marks
Student Modules Mark
D18 IMA Rank
Variable
D18 Individual Modules Average
D18 GSA Modules Average

Appendix 13. Ethical permission for the study

From: Sheena Smith <Sheena.Smith@durham.ac.uk>
Sent: 16 April 2007 10:41
To: richard.almond@cem.dur.ac.uk; Peter Tymms; Ma.secretaries
Subject: Ethical approval: Richard Almond
Attachments: ATTACHMENT_OR_CONTENT_BLOCKING_Sheena.Smith.vcf.TXT

Dear Richard
I am pleased to inform you that the Ethics Committee of the School of Education has now approved your application for ethical approval in respect of 'The effect that group summative assessment marking has on undergraduate course marks compared to individual summative marking'.

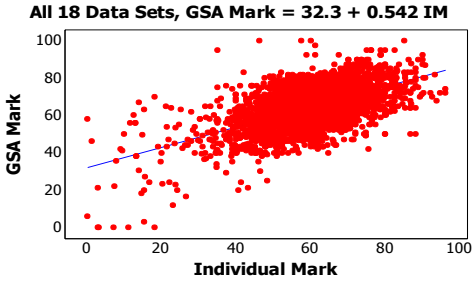
May I take this opportunity to wish you all the very best with your research.

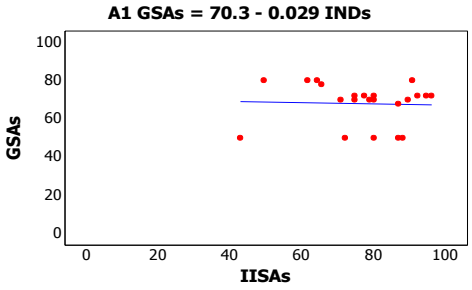
--
Sheena Smith
Durham University
School of Education

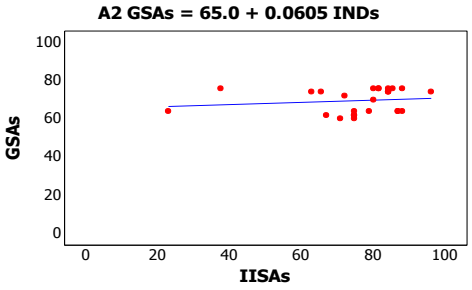
Tel: 0191 334 8403
Fax: 0191 334 8311

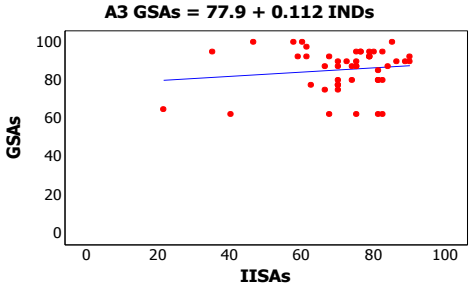
http://www.dur.ac.uk/education/

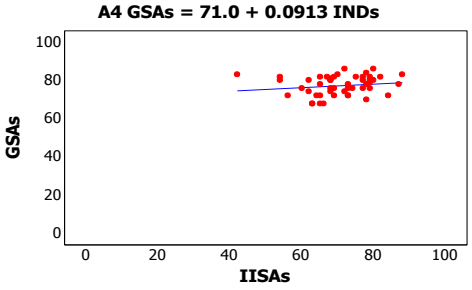
Appendix 14. Nineteen single regression-line scatterplots

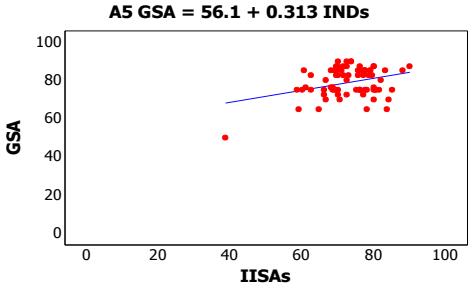


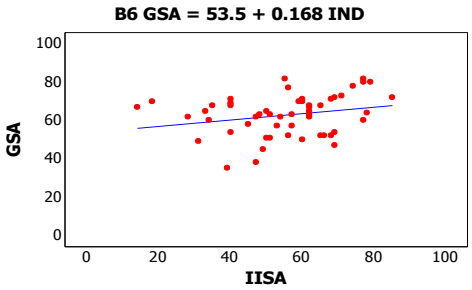


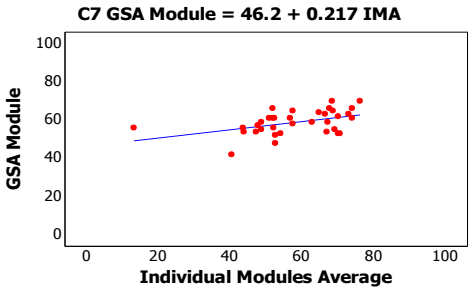


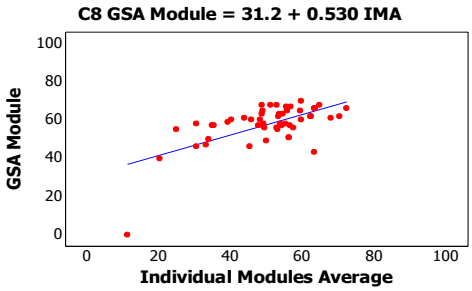


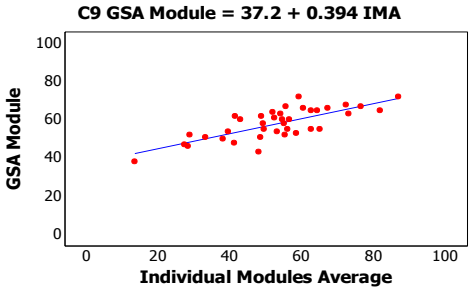


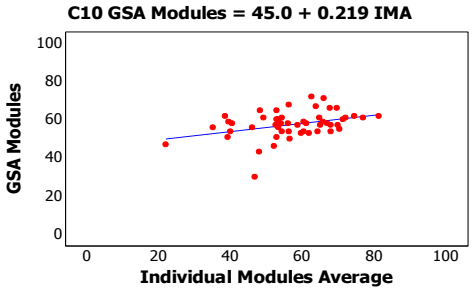


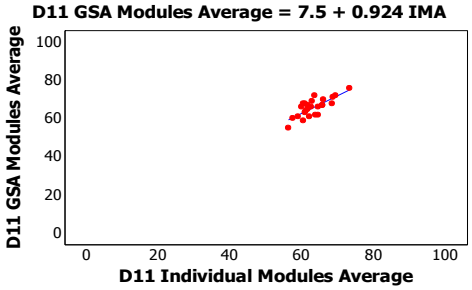


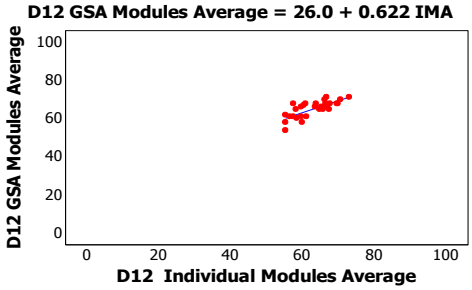

C7 GSA Module = 46.2 + 0.217 IMA
GSA Module
Individual Modules Average

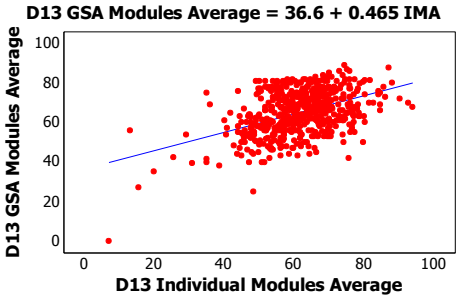

C8 GSA Module = 31.2 + 0.530 IMA
GSA Module
Individual Modules Average

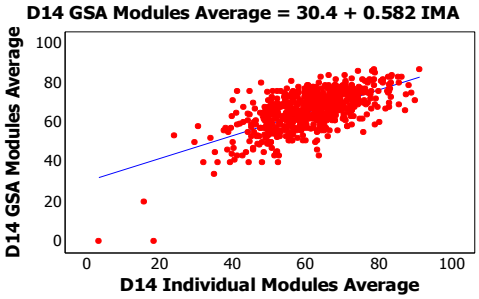

C9 GSA Module = 37.2 + 0.394 IMA
GSA Module
Individual Modules Average

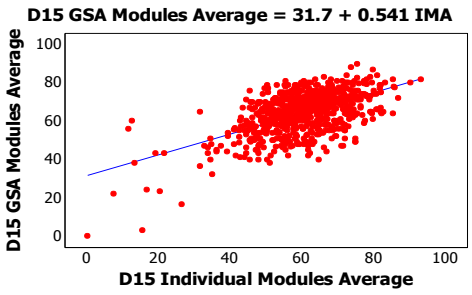

C10 GSA Modules = 45.0 + 0.219 IMA
GSA Modules
Individual Modules Average

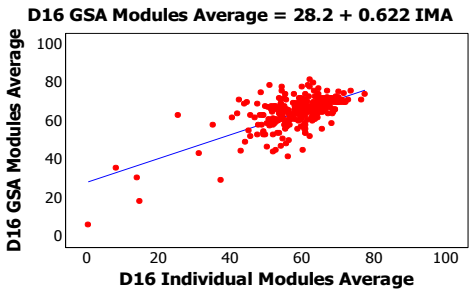

D11 GSA Modules Average = 7.5 + 0.924 IMA
D11 GSA Modules Average
D11 Individual Modules Average

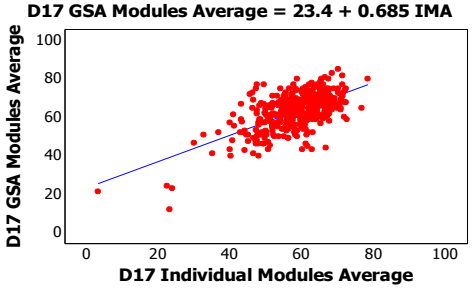

D12 GSA Modules Average = 26.0 + 0.622 IMA
D12 GSA Modules Average
D12 Individual Modules Average

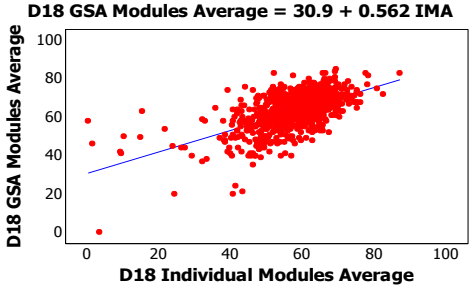

D13 GSA Modules Average = 36.6 + 0.465 IMA
D13 GSA Modules Average
D13 Individual Modules Average


D14 GSA Modules Average = 30.4 + 0.582 IMA
D14 GSA Modules Average
D14 Individual Modules Average


D15 GSA Modules Average = 31.7 + 0.541 IMA
D15 GSA Modules Average
D15 Individual Modules Average


D16 GSA Modules Average = 28.2 + 0.622 IMA
D16 GSA Modules Average
D16 Individual Modules Average


D17 GSA Modules Average = 23.4 + 0.685 IMA
D17 GSA Modules Average
D17 Individual Modules Average


D18 GSA Modules Average = 30.9 + 0.562 IMA
D18 GSA Modules Average
D18 Individual Modules Average

Appendix 15. Summary of marks from 18 data sets

Data ID	n	IISA mean	GSA mean	mGSA-mIISA	IISA sd	GSA sd	IISA SD - GSA SD	GSA (m+sd)
A1	22	77.0	68.1	-8.9	14.1	10.7	3.3	78.8
A2	26	75.6	69.5	-6.0	15.6	6.2	9.5	75.7
A3	48	71.5	85.8	14.4	13.9	11.0	2.8	96.9
A4	49	70.6	77.4	6.9	9.0	5.0	4.0	82.4
A5	74	73.0	79.0	5.9	8.0	7.5	0.5	86.5
B6	53	55.2	62.8	7.6	15.8	10.7	5.1	73.4
C7	35	57.9	58.8	0.9	12.9	6.1	6.8	64.9
C8	52	50.2	57.8	7.6	12.6	10.6	2.0	68.4
C9	38	52.8	58.0	5.2	15.5	8.0	7.4	66.0
C10	52	57.4	57.5	0.1	11.7	7.0	4.8	64.5
D11	25	63.1	65.8	2.7	3.9	4.7	-0.8	70.5
D12	31	62.9	65.1	2.2	5.0	4.1	0.9	69.2
D13	582	61.4	65.1	3.7	10.8	11.0	-0.2	76.1
D14	658	61.1	66.0	4.9	11.1	10.0	1.1	76.0
D15	751	60.3	64.3	4.0	11.5	10.9	0.6	75.2
D16	285	58.3	64.5	6.2	9.5	8.8	0.7	73.3
D17	445	58.2	63.3	5.1	8.4	9.2	-0.8	72.5
D18	844	57.4	63.2	5.8	9.8	9.4	0.4	72.6
Mean	226.1	62.4	66.2	3.8	11.1	8.4	2.7	74.6
Total:	4070							

Column GSA(m+sd) is the total of the GSA mean plus one standard deviation. This is the point below which, lie 84% of the data because, under a normal curve, 34.13 % of data lie between the mean and plus one standard deviation, i.e. 50 % + 34% = 84%. It allows assessment of the proximity of the raw data mark to the 100-mark ceiling, and therefore the potential to confound any statistical tests on the raw data based on the assumption of Normality. This value for the A3 data is almost 97. There is clearly something different about this module.

Appendix 16. Dual-lines regression models summary

Data Set	Sample size n	IISA Intercept	IISA Slope	GSA Intercept	GSA Slope	X axis bisect	x % of Range	x% without D11&17
A1	22	53.3	2.07	69.7	-0.14	7.43	34	34
A2	26	52.7	1.69	66.8	0.20	9.48	36	36
A3	48	49.9	0.88	85.1	0.03	41.51	86	86
A4	49	55.5	0.60	75.1	0.09	38.58	79	79
A5	74	60.2	0.34	76.7	0.06	57.96	78	78
B6	53	28.4	0.99	57.3	0.20	36.54	69	69
C7	35	37.0	1.17	53.8	0.28	18.81	54	54
C8	53	29.6	0.76	49.5	0.31	43.83	83	84
C9	38	26.7	1.34	47.7	0.53	25.80	68	68
C10	52	37.3	0.75	53.0	0.17	27.16	52	52
D11	25	56.4	0.51	59.9	0.46	61.40	246	*
D12	31	54.2	0.54	59.7	0.34	26.70	86	86
D13	582	43.8	0.06	57.5	0.03	398.26	68	68
D14	658	42.6	0.06	55.9	0.03	521.57	79	79
D15	751	41.8	0.05	55.3	0.02	535.71	71	71
D16	285	44.3	0.10	56.9	0.05	280.62	98	98
D17	445	44.8	0.06	54.9	0.04	448.89	101	*
D18	844	42.2	0.04	54.3	0.02	812.08	96	96
mean							82.5	71.3

*(IISA and GSA regression lines bisect outside the range of marks, i.e. beyond the 100 maximum possible score).

Appendix 17. Comparison of MLM raw and normalised data

A Comparison of 18 Data Sets MLwiN Generated Raw and Normalised Data Slopes Variances and Ranks. (See section 7.5)

	Raw Data Slopes		normalised Data Slopes	
Data Set	Variance	Rank	Variance	Rank
1	-0.296	1	-0.283	1
2	-0.168	3	-0.182	3
3	-0.268	2	-0.235	2
4	-0.161	4	-0.170	4
5	-0.091	5	-0.039	8
6	-0.066	6	-0.072	6
7	-0.028	8	-0.045	7
8	0.025	9	0.021	9
9	0.034	10	0.050	11
10	-0.053	7	-0.082	5
11	0.096	13	0.098	13
12	0.082	12	0.078	12
13	0.046	11	0.046	10
14	0.161	16	0.171	16
15	0.098	14	0.102	14
16	0.146	15	0.134	15
17	0.233	18	0.207	18
18	0.211	17	0.202	17

Appendix 18. Not all groups are teams: How to tell the difference

	Working group	Team
1	Strong, clearly focused leader	Shared leadership roles
2	Individual accountability	Individual and mutual accountability
3	The group's purpose is the same as the organizational mission	Specific team purpose that the team itself delivers
4	Individual work-products	Collective work-products
5	Runs efficient meetings	Encourages open-ended discussions and active problem-solving meetings
6	Measures its effectiveness indirectly by its influence on others (e.g. financial performance of the business)	Measures performance directly by assessing collective work-products
7	Discusses, decides, and delegates	Discusses, decides, and does real work together.
	(Katzenbach and Smith 1993)	

Appendix 19. Students' comments on their group work experiences

Table 1: Students' positive comments regarding their group work experiences

Group work is interesting

We are able to work on larger projects especially those involving fieldwork and projects with real businesses

It exposes us to a greater variety of viewpoints and ideas and enables ideas to be critically evaluated on the spot

It generates synergy

We can draw on the special skills of group members to the benefit of the whole group

It helps us to learn to work together and improves communication skills

It reflects real life - you learn how to deal with conflict and not allow it to interfere with the work that needs to be done

Table2: Students' negative comments regarding their group work experiences

There are problems with time management and learning and dealing with the mechanics of group operation

The logistics of getting together

Not learning to the full extent those parts for which other group members are responsible

Distributing the workload evenly

Free riders benefiting from the group effort and marks being lowered because of some members not meeting their obligations

Personality conflicts impeding the effective operation of the group

(Morris and Hayes 1997)